

Complacency about misconduct

Plagiarism is a serious sin, but universities and journals do not always respond appropriately. A case reported this week suggests that some in the physical sciences have yet to appreciate the threat to confidence in science.

Researchers who choose to deceive their colleagues can adopt a variety of strategies. Details of a physicist who employed a low-risk, low-reward approach — the plagiarizing of papers from foreign-language journals — came to light this week. And although the fraud produced papers that have amassed few citations, the response of the researcher's peers is cause for concern.

Yung Park was certainly no Jan Hendrik Schön. The latter was a high-profile young physicist at the prestigious Bell Laboratories in Murray Hill, New Jersey. When suspicions about Schön surfaced in May 2002, his employers went public, launched an inquiry, and helped to ensure that journals were alerted promptly so that their readers were soon aware of the problem. Schön was found to have fabricated and falsified data in 25 papers — and physicists agreed that they needed to take a hard look at how they handled cases of misconduct.

Responses to Park's misconduct have been less commendable. Between 1995 and 2002, while at the Korea Advanced Institute of Science and Technology (KAIST) in Daejeon, and then as a visiting scientist at the University of Cambridge, UK, Park published about 80 papers in materials science. At least eight of these are plagiarisms, most of them almost identical copies of papers originally published in Russian. When confronted by the evidence in April 2002, Park left Cambridge and his whereabouts are now unknown (see page 3).

Unfortunately for researchers who work in Park's field — the study of electronic properties of materials — Park's former colleagues at Cambridge failed to notify journal editors of some of the plagiarisms that they uncovered. Some editors who were informed by KAIST took no action. As a result, at least four known

plagiarisms remain available in journals.

Four papers, all of low impact, may not seem much cause for concern. This indeed seems to have been the attitude of many of those involved. But although journal editors and researchers contacted by *Nature* insist on the need to combat plagiarism, their lack of action indicates that parts of the physics community feel that it is not their responsibility to deal with misconduct.

This is odd, as simple steps can help put right plagiarism of the type practised by Park. The journal *Solid State Communications* published a paper by Park that is now known to have been plagiarized. The editors have attached a note to the online version of the paper, marking it as a plagiarism and directing readers to the original work. The journal has also published a note in print. Together, these measures should minimize any loss of citations that the author of the original paper might suffer, and make it clear that Park is guilty of misconduct.

Research institutions have responsibilities too. As a visiting scientist, Park was not paid by Cambridge. But he worked in the university's labs and published under its affiliation. The university therefore has a duty to investigate all of the 40 or so papers published by Park while he was there, and to make the results of that assessment available to journal editors and the research community.

The effect of not taking such steps is clear. Authors of the original papers lose out on citations, and the chance that Park may dupe another employer increases. And when the misconduct and a lack of response are made public, public confidence in science takes another small knock. Park's plagiarism will not rock the physics community in the way that Schön's fraud did, but that is no excuse for acting as if it didn't happen at all. ■

Nature's twentieth-century highs

Celebrating a book about discoveries that changed science.

Every scientist, like Newton, stands on the shoulders of giants, but few stop to look directly at what those giants accomplished. The history and the documentation of all but the most recent of past discoveries is of concern only to historians, sociologists and scientists who are no longer actively engaged at the front line.

That, at least, is a view held by most young researchers. But a new book warrants their attention and that of anyone else interested in science: *A Century of Nature: Twenty-one Discoveries that Changed Science and the World* (University of Chicago Press).

Here we must declare an interest: the book is wholly devoted to this journal's content, and the editors of the book are, or were, colleagues. Laura Garwin was our North American editor and Tim Lincoln is our News and Views editor. But according to Freeman Dyson, "this book provides a much more solid basis for scientific literacy than the many popular books that are devoted to the latest scientific fad", and Jared Diamond and Steven Weinberg have celebrated its coverage across the disciplines. So the book's editors

can be proud of it and we can be grateful to them and promote it with a clear conscience.

Some of the original papers almost leap off the page. The report in 1925 by Raymond Dart of his discovery of *Australopithecus africanus*, the first fossil link between apes and man, unabashedly presents a personal narrative. James Chadwick's report in 1932 of the discovery of the neutron has an almost casual brevity and informality about it that today's authors can only envy. In other cases it is the accompanying essays, many written by those working close to the original research, that bring the papers to life.

But whatever the topic — plate tectonics, extrasolar planets, T-cell immunology, the ozone hole, the generation of animal body plans, cloning — the essays also entice the reader into a far greater appreciation of the work than can be obtained when it is transmitted through textbooks. This collection of seminal papers shows that, like much-reproduced great paintings, however one may have been impressed at secondhand, there is something uniquely inspiring about the originals. ■





Under threat
TV programme could
cost German research
lab its licence
p4



Pet project
Argentina proposes
export plan for trade
in wild parrots
p4



Wild thing
Comet-chasing probe
prepares for dust
collection
p6



Bjørn again
Danish committee
asked to revisit
Lomborg's book
p7

Plagiarism in Cambridge physics lab prompts calls for guidelines

Jim Giles, London

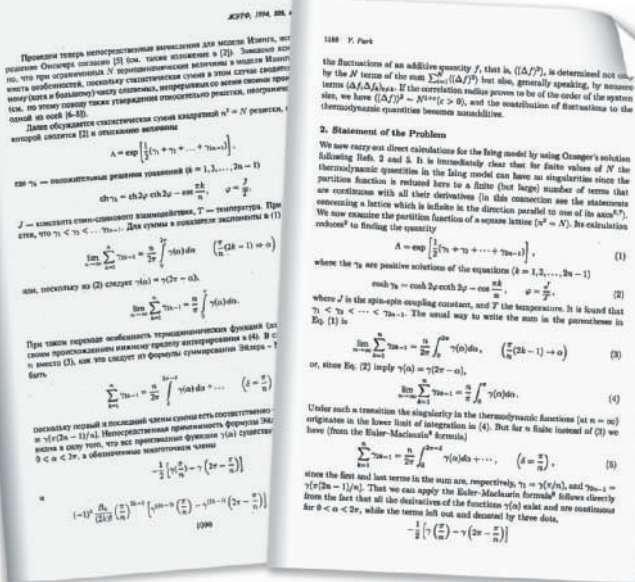
A string of plagiarized papers remains unaltered in the literature, after researchers and journal editors ignored warning signs of the problem, an investigation by *Nature* has found.

At least eight papers written between 1997 and 2001 by Yung Park, a materials scientist who worked at the University of Cambridge, UK, and the Korea Advanced Institute of Science and Technology (KAIST) in Daejeon, are plagiarized, according to documents examined by *Nature*. Park has also published at least two pairs of papers with significant overlap in separate journals.

Four of the plagiarized papers have been removed from online journals, or marked as plagiarisms, following investigations into Park's publications by his former colleagues in Korea and by researchers whom he plagiarized. But others remain available either because researchers failed to pass on evidence of plagiarism to journal editors, or because editors were informed and took no action. The status of many of Park's other papers — he published about 80 between 1995 and 2002 — has not been fully investigated.

Park came to Cambridge in 1997, where he worked in the lab of materials scientist Kevin Knowles as a visiting scientist under a fellowship from the British Council. Doubts about Park's work were first raised in April 2002 when Bagautdin Bagautdinov, a materials scientist based at the Himeji Institute of Technology in Japan, sent Knowles evidence that Park¹ had plagiarized one of his papers².

Together with Derek Fray, head of the materials science and metallurgy department at Cambridge, Knowles investigated the allegation. He says that Park tried to show that the work was original, but was unable to provide convincing evidence. As a visiting scientist who was not getting a salary or research funding from Cambridge, Park was asked to leave immediately. He is believed to have returned to Korea, but attempts by



Twins: the paper on the right closely adheres to the Russian original (left).

Nature to trace him have failed.

Following Park's departure, other examples of plagiarism in his papers have been identified. For instance, Feodor Borodich, a materials scientist then at the University of Liverpool, spotted three cases and sent details to the editors of the 19 journals that Park had published with while at Cambridge. KAIST researchers investigated papers published by Park under the institute's affiliation and e-mailed six journal editors in June 2002 informing them of one further case of plagiarism and three sets of 'overlapping' papers.

But some other enquiries were not followed up. Knowles spotted three more cases of plagiarism, yet never contacted the editors of the journals involved. One of these papers has since been marked online as plagiarized after it was brought to the attention of the journal in question, but the other two remain available^{3,4}.

Several journals also failed to act when shown evidence of possible misconduct. The *Journal of the American Ceramic Society*, for example, which published a paper by Park that shows significant overlap with a previously published paper, says that it received an e-mail from KAIST but only began inves-

tigating the matter when contacted by *Nature*. The *Journal of Physics D: Applied Physics*, which published a plagiarized paper by Park, confirmed when contacted that it had received the KAIST e-mail but took no action.

Knowles, who was a co-author on seven of Park's papers that have not come under suspicion, says that in retrospect it was an error for him not to inform editors of the cases of plagiarism that he identified. He says that he is now more aware of misconduct issues, and recently notified journal editors of a pair of duplicate publications not involving Park that he stumbled across.

But the fact that Park was a visiting researcher seems to have influenced Cambridge's approach to the problem. "We hoped it would go away," Fray said when first contacted by *Nature*. "The person had nothing to do with us. It would have been different if he had been employed by us."

Journal editors who did investigate allegations against Park say that they are disappointed by the responses of other journals. Martin Blume, editor-in-chief at the American Physical Society in Ridge, New York, checked a paper by Park in *Physical Review B* after being contacted by Borodich. He found no evidence that it was fraudulent. "But when you learn about possible misconduct, you've got to do something," he says.

Editors say that the incident highlights the lack of guidelines for handling plagiarism. The need for a code was discussed last October at a meeting run by the International Union of Pure and Applied Physics in London. Blume, one of the meeting's organizers, says that the union is developing draft guidelines, which should be ready by the autumn.

1. Park, Y. *Europhys. Lett.* **52**, 557 (2000).
2. Bagautdinov, B. Sh. & Shmyt'ko, I. M. *JETP Lett.* **59**, 182–186 (1994).
3. Park, Y. *Solid State Commun.* **115**, 281–285 (2000).
4. Park, Y. *Int'l J. Mod. Phys. B* **14**, 1187–1194 (2000).

Primate lab faces closure threat over mistreatment charge

Quirin Schiermeier, Munich

Germany's largest primate laboratory is in danger of losing its licence after a television exposé alleged mistreatment of monkeys there.

Last spring, the British Union for the Abolition of Vivisection sent film-maker and investigative journalist Friedrich Mülln undercover into the Münster laboratory of Covance, a multinational



TV footage of animals at Covance's laboratory has shocked the German public.

company that carries out contract research for the pharmaceutical industry. Over five months he secretly filmed staff seemingly tormenting macaques and rhesus monkeys.

The footage, shown last month on German public television, shows animal keepers apparently dancing with semi-anaesthetized macaques, shaking the monkeys' heads in time to music.

Bärbel Höhn, environment minister of North Rhine-Westphalia, has asked public prosecutors to investigate the "obvious unreliability" of Covance's staff. If the programme's accusations are justified, the laboratory's licence to keep primates will be withdrawn, she says.

"We are convinced that our staff have not violated animal-protection laws," says Friedhelm Vogel, managing director of the Münster lab. "The scenes are authentic. However, they have been edited and commentary added such that the impression arises that monkeys were systematically mistreated. This is not true." Covance has said that it will cooperate fully with the authorities and take any action required.

"It does appear as if animals were at least treated in a degrading way," says Ivar Aune, spokesman for the German research lobby group Society for Health and Research. But he adds that "monkeys are dangerous animals and precautions have to be taken".

Ecologists hit out at plan to export Argentinian parrots

Rex Dalton, San Diego

An application from Argentina to export parrots as pets to the United States has sent feathers flying among ecologists, some of whom say it would deplete populations in the wild.

If the application is accepted, Argentina will become the first country to export wild birds for the US pet trade since a tough law on imports was enacted in 1992 to prevent species being wiped out by commercial trappers.

Government biologists in Argentina have fashioned "a sustainable-use management plan" that they say will allow limited collecting of the blue-fronted Amazon parrot, *Amazona aestiva* — with part of the money from sales going towards maintaining the parrots' threatened forest habitat.

Such plans are seen as a way of preventing the extinction of animals and plants in countries where forest is being clear-cut. Argentina — South America's biggest exporter of wild birds — sends blue-fronted Amazons to Europe and elsewhere under its plan.

But Argentina's proposal, which the US Fish and Wildlife Service (FWS) sought comments on last August, has been criticized by about 100 ecologists and 30 conservation organizations around the world, who have lodged objections to it with the FWS.

The critics claim that there has been insufficient scientific study of the implications of harvesting blue-fronted Amazons, that inadequate resources are put aside to monitor the collecting and that the plan fails to address the possibility of exported parrots escaping to the wild and transmitting exotic diseases to native birds.

The blue-fronted Amazon is not endangered or threatened. But opponents of the application claim it could set off a rush of applications to export other species to the United States, potentially depleting wild-bird stocks in South America, Africa and Asia. "We are very concerned about the precedent this will set," says James Gilardi, director of the World Parrot Trust based in Cornwall, UK.

The Argentinian proposal is defended by tropical ecologist Susan Lieberman, director of environmental group the WWF's species programme in Surrey, UK. As a former official of the FWS, she was involved in drafting the 1992 legislation.

The Argentinian plan "is not a perfect situation", she says. "But they have made a valiant effort. I think the plan is tenable." If it does not seem to be working, the plan can be suspended when it comes up for renewal in three years' time.

"The parrots are losing out to agriculture that removes habitat," says Ricardo Banchs, a biologist with Argentina's Sustainable Development and Environmental Secretariat. "There is no easy solution. We believe this is the only way to solve the problem."

Ecologist Enrique Bucher, an authority on the blue-fronted Amazon at the National University of Cordova in Argentina, is against the proposal. "We don't have a scientific way to manage these things," he says.

Michael Kreger of the FWS says that the agency is reviewing the comments received. "We want the Argentines to do this based on the best science," he says. "They aren't doing this to exploit the resource." A FWS decision is expected in the next few months.



Sales talk: parrots are bought and sold in the street markets of South America.

Mars satellite flies into hunt for lost Beagle 2

Declan Butler, Paris

Since Christmas Day, British mission controllers have listened in vain for a signal from their Beagle 2 Mars lander. Now the orbiter that carried the probe to the red planet is about to join in the search.

On 4 January, the Mars Express craft, operated by the European Space Agency (ESA), will begin sweeping the surface of Mars for radio signals from Beagle 2. The mother ship represents the best chance of finding out whether the lander survived its descent to the martian surface.

Beagle 2 was jettisoned by Mars Express on 19 December, a camera on board the ESA craft photographing the pocket-watch-shaped lander falling away on the final 3 million kilometres of its 400-million-km journey. It should have entered the martian atmosphere at around 20,000 km per hour at 2.45 am GMT on 25 December, before landing in the crater of Isidis Planitia about six minutes later, slowed to some 60 km per hour by parachutes and cushioned by airbags.

Just under three hours after that, mission controllers had hoped that Beagle 2's calling sign — a nine-note tune composed by the pop band Blur — would be picked up by NASA's Mars Odyssey craft, which has been orbiting the red planet since 2001. Over the days that followed, Mars Odyssey repeatedly tried and failed to make contact with the lander. The 76-metre Lovell radio telescope at the Jodrell Bank observatory in Cheshire, UK, and a 45-metre dish at Stanford University in California also drew a blank.

For the team behind Beagle 2, led by Colin

Pillinger of the Open University in Milton Keynes, Beagle 2's silence has been frustrating. "This is a bit disappointing, but it's not the end of the world," says Pillinger. Mars Express and Beagle 2 were designed to talk to each other, so the addition of the mother ship to the search party should give the best chance of detecting the lander.

Mars Express itself slipped into orbit in the middle of Christmas morning — an impressive result for ESA's first planetary mission. This weekend, mission controllers will fire the orbiter's engine for three minutes to nudge it from equatorial orbit into its final polar orbit, where it can begin operating on 4 January.

This new orbit will allow frequent passes over Beagle 2's landing area. At around the same time, the lander, which currently only broadcasts at set times, is programmed to switch to an emergency mode in which it will broadcast continuously during daylight hours.

Despite Beagle 2's high media profile — a product of Pillinger's gift for publicity (see *Nature* 423, 476; 2003) — Mars Express is the



The Beagle is stranded: Colin Pillinger has tried in vain to contact his Mars lander since its scheduled touchdown on Christmas Day.

main part of ESA's mission. It carries a high-resolution camera to map the planet, a near-infrared spectrometer to assess surface mineralogy, radar to search for subsurface water down to a depth of 5 km, and instruments to study the planet's upper atmosphere and its interaction with the solar wind.

While the search for Beagle 2 continues, two NASA exploration rovers are also on their way to Mars (see *Nature* 423, 477; 2003). The first, Spirit, is also due to land on 4 January; the second, Opportunity, will touch down on 25 January.

♦ www.beagle2.com

♦ www.esa.int/mars

Beef blockade greets first mad cow in United States

Rex Dalton and Erika Check

For the United States Department of Agriculture (USDA), it was about worst the Christmas present imaginable: the country's first case of mad cow disease. Nations around the world responded by banning the import of US beef.

The diagnosis of bovine spongiform encephalopathy (BSE) — in a 'downer' cow, unable to walk yet still sent for slaughter and human consumption — was confirmed on 25 December. Two days later, the USDA tentatively traced the animal to a herd imported from Alberta, Canada, which announced North America's first 'home-grown' case of BSE last May (see *Nature* 423, 467; 2003). A previous Canadian case, in 1993, occurred in a cow imported from Britain.

The new US case is from a farm in Mabton, Washington, about 200 kilometres

southeast of Seattle. Efforts are being made to recall potentially contaminated meat products. Even though the cow may not have been infected in the United States, the case has raised questions about the adequacy of measures designed to keep BSE out of the US cattle herd — and to protect people from infection with the brain-wasting disease. In particular, scientists are calling for an expansion of the BSE-testing programme for slaughtered cattle.

The animal was probably infected by eating feed contaminated with the 'prion' proteins that are thought to cause the disease. Both the United States and Canada have banned the feeding to cattle of meat and bone meal from sheep and cows, although the meal can still be fed to other livestock. Experience in Britain, which introduced a similar ban in 1988, suggests that such bans are difficult to

enforce unless meat and bone meal from ruminants is completely excluded from animal feed.

Experts also say that too few cattle are being tested for BSE in the United States to be sure that infected animals are not entering the human food supply. There is no routine testing for the 30-million-plus cattle slaughtered each year — and only about 10% of the 200,000 or so downer cattle slaughtered in 2003 were tested, even though an inability to walk is a known symptom of BSE. "We should be testing more animals," argues Frederick Murphy, a veterinary scientist at the University of California, Davis.

The case is also likely to lead to a revival of attempts to exclude all downer cows from the human food supply. Such efforts have previously been defeated after lobbying from agricultural interests.

UK medical-research chief angered by snub in Queen's honours list

London The head of Britain's Medical Research Council, Colin Blakemore, has threatened to resign after a newspaper report claimed that he was denied a knighthood because of his vocal defence of research on animals.

According to papers leaked last month to *The Sunday Times*, Blakemore was refused the honour because civil servants considered his public statements on animal experimentation to be too controversial. Blakemore has received letter-bombs from animal-rights activists in the past.

In the wake of the report, Blakemore threatened to resign unless the government reaffirmed its support for animal studies and acknowledged the need to defend the researchers involved. Science minister David Sainsbury announced on 22 December that the government "admires and fully supports those on the front line who have stood up to animal-rights extremists".

Sainsbury also stressed that the decision was made by civil servants and does not reflect government policy. Blakemore said that he was considering his response to Sainsbury's comments.



Stardust has a date this week with Comet Wild 2.

NASA probe closes in on comet to catch stardust

Washington Almost five years after it was launched, NASA's Stardust spacecraft is finally closing in on its quarry. On 2 January the probe is scheduled to fly through the dust surrounding Comet Wild 2, aiming to snag dust particles to bring to Earth in 2006. The project is the first attempt to capture cometary material for study on the ground.

Stardust's collecting device will use a foam-like gel to attempt to trap the interstellar dust grains without shattering them as the probe zips through the cloud at six kilometres per second. Cameras on the spacecraft should also return pictures

showing Wild 2's nucleus in unprecedented detail. The European Space Agency's Giotto mission photographed Halley's comet at close range in 1986, but Stardust's planned fly-by will be much closer, and the comet's dust cloud is not expected to be as thick.

Chinese test results spark fresh SARS fear

Tokyo SARS may have returned to China's Guangdong province. As *Nature* went to press, World Health Organization officials were scrambling to investigate test results from three labs in China, which suggested that a 32-year-old television producer from the region had been infected by the virus that causes severe acute respiratory syndrome.

Confusion surrounded the diagnosis, however, as not all of the tests conducted on samples from the patient yielded positive results. His temperature also returned to normal faster than would be expected.

The suspected Chinese case follows an incident in neighbouring Taiwan, in which a researcher at the Institute of Preventive Medicine of Taiwan National University, near Taipei, infected himself after handling samples without wearing protective gloves. This was the second instance of laboratory contamination, following a case in Singapore in September.

The cases raise questions about lab safety — but public-health experts are much more alarmed about the suggestion that SARS has returned to Guangdong, which was the epicentre of last year's outbreak.

Panel ordered to revisit verdict on Lomborg book

London An investigation into the scientific honesty of Bjørn Lomborg, the author of a popular book that attacks many widely held environmental beliefs, has been re-opened by the Danish government.

The Danish Committees on Scientific Dishonesty, which studied Lomborg's book, *The Skeptical Environmentalist*, said in January 2003 that the work was, "objectively speaking, deemed to fall within the concept of scientific dishonesty" (see *Nature* 421,

201; 2003). The assessment was widely criticized, partly because it relied on existing critiques in popular science magazines.

The Danish science ministry decided to review the committees' decision, and on 16 December asked the committees to repeat the investigation. The ministry said that the committees had failed to document exactly why Lomborg was guilty.

The same government appointed Lomborg to his current position as head of the Environmental Assessment Institute in Copenhagen in 2002.

Prospects dimmed for glowing cell markers

San Diego 'Quantum dots' — tiny particles of metal or semiconductor hailed as a groundbreaking new material for labelling cells — could turn out to be toxic to the very cells they are being used to tag.

Nanometre-sized dots of cadmium selenide emit visible light when exposed to ultraviolet radiation, making them ideal for labelling cells (see *Nature* 413, 450–452; 2001). But Austin Derfus and his colleagues at the University of California, San Diego, have found that the dots damage cells under certain conditions (A. M. Derfus *et al.* *Nano Lett.* doi:10.1021/nl0347334).

Oxidation of the dots' surface, by

exposure to ultraviolet light, for example, leads to the release of toxic cadmium ions. Although the researchers say that surface coatings can help to reduce toxicity, they warn that this may not prevent cell damage in longer cell-labelling experiments.

Fusion reactor to stay homeless until New Year

Washington The decision on where to host ITER — a US\$5-billion experimental fusion reactor — has been postponed until February.

Some physicists expected a final site to be selected at a conference on 20 December near Washington DC. But it became apparent at the meeting that the two candidate sites, in France and Japan, were in for a close fight.

The United States supported the Japanese site in the northern province of Rokkasho, whereas Russia and China supported the European Union's proposed venue in Cadarache, France. South Korea eventually gave its support to the Japanese site, leaving the six project members equally split.

Negotiators said that they will continue to work towards an agreement that will satisfy all partners. One proposed solution is to construct the reactor at one site, while building a computer-modelling centre, an advanced materials laboratory and a remote control-room for the project at the other.

PA/EPA



Bjørn Lomborg: dishonesty claims criticized.

Organ failure

Across continental Europe, historical instruments are falling silent, muted by a new and mysterious form of corrosion. Tom Clarke speaks to the chemical detectives who are striving to protect our musical heritage.

Since 1467, the pure sound of the Stellan organ has filled the parish church of St Jakobi's in Lübeck, Germany. Named after the great organ builder Friedrich Stellwagen, who renovated it in the seventeenth century, aficionados rate the Lübeck organ as among the world's best for performing Renaissance and early baroque music. But in 1992, its largest pipes began to lose their voice. Close inspection revealed the problem: air was escaping through tiny holes that had appeared in the metal.

News of the renowned organ's affliction spread quickly among Europe's close-knit community of organ builders, players and enthusiasts. And the word in the organ loft was that Lübeck's Stellwagen was not alone. Pipes in many instruments dating from the fifteenth, sixteenth and seventeenth centuries from churches in Belgium, France, Italy, the Netherlands and Portugal are similarly rotting away — those of the organ in Bordeaux Cathedral are in danger of collapsing under their own weight.

The pipes of ancient organs are made from lead, which is known to corrode. But the symptoms seen in the Lübeck Stellwagen and its fellow sufferers — the sudden appearance of white chalky residue on a pipe's interior that eventually works its way through to the outside — hadn't been encountered previously. "We really think that we're dealing with something new," says Carl Johan Bergsten, a research engineer at the Organ Art Center at Gothenburg University, Sweden. Bergsten views the instruments as both musical and technological icons. "They mirror the technical achievements of their age," he says. "An organ was the seventeenth century equivalent of the PC."

Following the lead

Last year, Bergsten assembled a team of metallurgists, chemists, organ makers and music historians to establish the cause and extent of the problem. It's early days, but the team's laboratory and field experiments have started to yield clues about the cause of the corrosion — and it seems that it may be a by-product of well-meaning attempts to restore the precious instruments.

Bergsten's Corrosion of Lead and Lead-Tin Alloys of Organ Pipes in Europe project — funded by the European Union and known by its acronym, COLLAPSE — is inviting organ builders, restorers and organists to send it details of corroded organs. Just last month,



Finely tuned: lead samples from organs are tested for corrosion under various atmospheric conditions.

for example, the Royal Conservatory in Brussels contacted Bergsten after realizing that its organ had fallen victim to lead corrosion.

Understanding the distribution of the problem might help to identify its cause. But the project's main detective thrust is a series of field and lab experiments. Bergsten's team has already selected seven corroded organs in Italy, Germany and the Netherlands and paired them with similar non-afflicted organs in similar churches in the same area. Instruments recording temperature and humidity have been installed in these churches, and samples of corroded and uncorroded metal from their organ pipes have been examined in the lab of Jan-Erik Svensson, an environmental inorganic

chemist at the Chalmers University of Technology in Gothenburg.

Organ enthusiasts weren't short of theories about the cause of the corrosion. One leading suspect was central heating. As well as increasing the temperature and humidity, heating systems also release carbon monoxide, which can corrode metals. But Svensson's analysis suggested an alternative cause. "We immediately found high concentrations of organic acids," he explains. The powdery white residue turned out to be lead hydroxycarbonate and lead hydroxyacetate, symptomatic of organic-acid corrosion. Sampling inside the corroded pipes immediately pointed to a culprit: high levels of acetic acid in the air blowing through them.



Holey pipework! Corrosion is causing the Stellwagen organ in St Jakobi's, Lübeck, to lose its voice.



Winded: many lead organ pipes in Europe, such as this one from the organ above, are disintegrating.

Further tests in the lab looked at other corrosive agents that might be involved. Common atmospheric pollutants such as sulphur dioxide, nitrogen oxides and ozone did little damage to samples from the pipes. But acetic acid, at concentrations similar to those found in organ pipes, was highly corrosive. "I've never seen such a good correlation between laboratory and field results," says Svensson.

Subsequent investigations revealed the likely source of the acid. New oak wood gives off high concentrations of acetic acid, and restoring an organ often entails rebuilding oak components in its bellows and wind chest — the box from which high-pressure air is delivered to the pipes. Most of the organs suffering from corrosion had been restored

in the recent past — the Lübeck Stellwagen was overhauled in the 1970s, for example.

But many old organs have had their wooden parts replaced several times without falling victim to corrosion. So why has the problem only started to emerge in recent years? This could be where the enthusiasts' hunch about central heating fits in. Svensson is eagerly awaiting the results of humidity and temperature monitoring; it could be that warmer church air is driving off more acetic acid from the new wood, he says.

Another key piece of evidence to emerge from the COLLAPSE study is that all the affected organs were built in the 'North German' style, and their lead pipes contain a small percentage of tin. This lead was first cast into

sheets on sand, and tin was added both to harden the pipes and to increase their lustre. But when most of the organs that are now corroding were made, tin was scarce and expensive, and so could only be used sparingly.

Metallurgist Carla Martini at the University of Bologna in Italy and her colleagues are probing the affected pipes' composition using atomic absorption and X-ray fluorescence spectroscopy. This has confirmed that the corrosion seems to occur only in pipes containing 1.5–2% tin. "These traces of tin seem to have a big influence on corrosion," Martini says.

Tinned goods

This may explain why most organ pipes in Britain seem to be immune to the phenomenon. "I've only seen two cases in my 30-year career," says John Norman, a London-based consultant who advises the Churches Conservation Trust on organ preservation. At the time the corroded pipes were made, the main source of Europe's tin was Cornwall. As a result, it was much cheaper in Britain than in continental Europe — so British organ pipes contain up to 20% tin.

Why corrosion should occur only in low-tin pipes remains a mystery. The metal itself isn't responsible, says Martini. Her optical- and electron-microscope analyses are revealing that the low tin-content pipes also contain large amounts of impurities such as copper, antimony and bismuth. These influence the pipes' microstructure — the size and arrangement of crystal-like pieces of metal separated by air spaces and impurities. Martini suspects that understanding how the impurities promote corrosion is likely to provide the key to the mystery. But after hundreds of years, it is difficult to know how the pipes' composition has changed over time. "There's a lot we may never understand," says Martini. "One thing you cannot emulate is time passing."

Lovers of organ music can only hope that the COLLAPSE project produces a cure for the corrosive condition before too much more time elapses. Bergsten and his colleagues aim to devise a way to treat the pipes chemically to prevent the reaction that is gnawing away at them. In the longer term, they hope to understand the relationship between the composition and manufacture of the old pipes, and their wonderful sound. Then it might be possible to replace badly corroded pipes with new ones, without compromising the instruments' distinctive voices.

Aficionados can only hope that the project makes rapid progress. For now, the Lübeck Stellwagen is still playable, says Lutz Jeddeck, pastor of St Jakobi's. "Works take on a special freshness due to the glorious and unmistakable tonal colour of the instrument," he enthuses. "But the holes get larger and larger." ■

Tom Clarke was until recently in *Nature's* online news team; he is now science reporter for Channel 4 News in London.

► www.goart.gu.se/collapse

All wired up

The ocean floor is being covered with remote-controlled observatories, letting oceanographers keep tabs on the sea without getting wet. Jon Copley investigates.

The ocean floor is an exciting place to visit, but you might not want to live there. Fortunately, those who want to keep an eye on what goes on beneath the waves now have an alternative. Instead of trying to colonize the sea floor, marine scientists are planning networks of unstaffed seafloor observatories across the globe. Connected by fibre-optic cables or linked to satellite buoys, this wet-world-web will let oceanographers make long-term measurements and run experiments in the depths without leaving their desks.

Back in the 1960s it seemed we were on the verge of living under the sea, gadding about in submersibles while pet porpoises fetched the morning papers. To show that people could live underwater, Jacques Cousteau famously developed the Conshelf habitats of basic undersea homes that could accommodate several people at depths of 10 to 100 metres, complete with a garage for a diving saucer. The US Navy SEALAB experiments, which actually used a porpoise to deliver mail from the surface, explored the limits of undersea living further. But the dream of colonizing the continental shelf proved costly and dangerous. Only one such project — the Aquarius habitat of the US National Oceanic and Atmospheric Administration (NOAA) — is still used as a research station, under 20 metres of water in the Florida Keys.

Nevertheless, a long-term window on the ocean depths has remained oceanographers' dream. The classic method of collecting oceanographic data by dropping instruments off the side of a ship comes with a persistent problem. It only provides a snapshot of the temperature, salinity and other properties of a single mass of water at a single point in time. Picking out trends from such pinpoints of data can be tricky. Some missions aim to sail the same track repeatedly, collecting the same kind of data each month. Experiments of this kind off the coasts of Hawaii and Bermuda have shown an intriguing fluctuation in the extent that these ocean regions act as a sink for atmospheric carbon dioxide^{1,2}. But divining the extent of, or the

reason for, those fluctuations is difficult without more data. Satellites, on the other hand, can provide a vast amount of nearly continuous data over great swathes of ocean. But they cannot pick out most of what happens beneath the surface. Ideally, marine scientists need a network of stations that give good coverage in both space and time.

Take the plunge

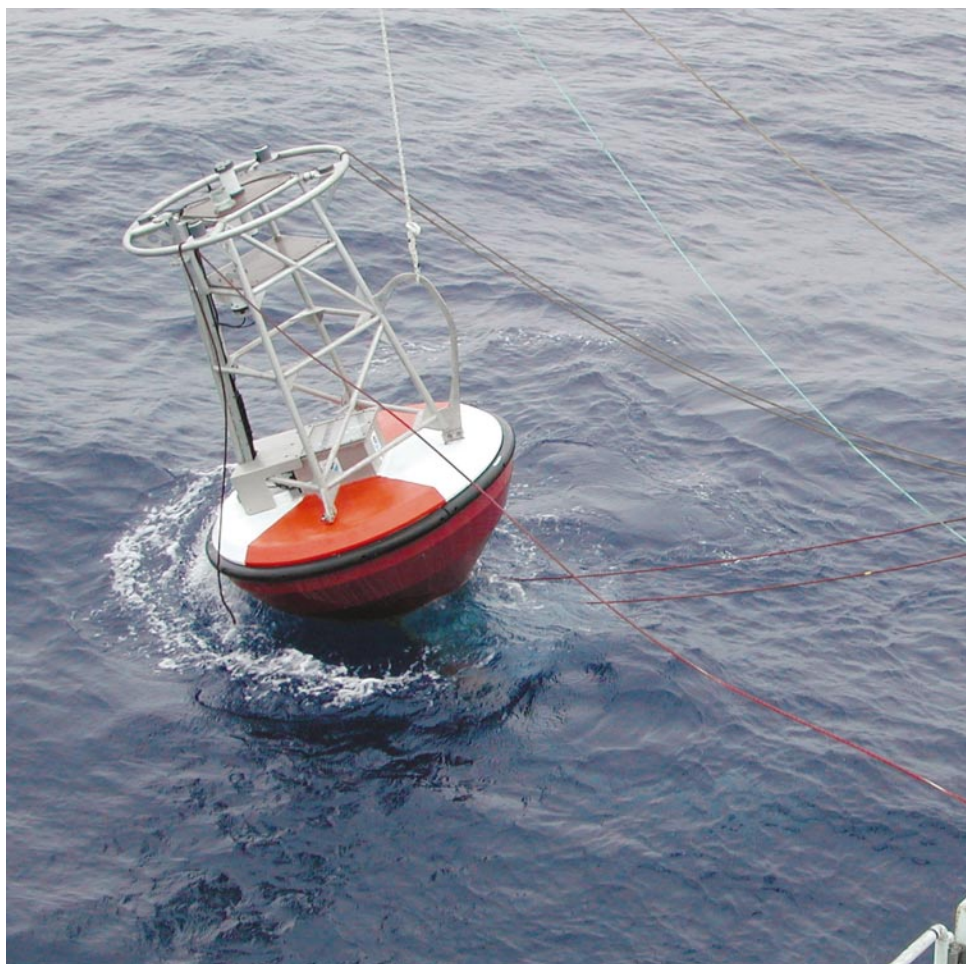
Underwater observatories can do just that. Early in January researchers will converge on San Juan in Puerto Rico to thrash out the priorities for future US efforts in this field. As part of a new Ocean Observatories Initiative, the National Science Foundation (NSF) has earmarked US\$200 million for infrastructure investment for both coastal and deep-ocean observatories. Meanwhile, researchers in Europe and Japan are also exploring ambitious plans for constellations of seafloor observatories. Over the next few decades it seems likely that a web of observatories will be spun out into the abyss.

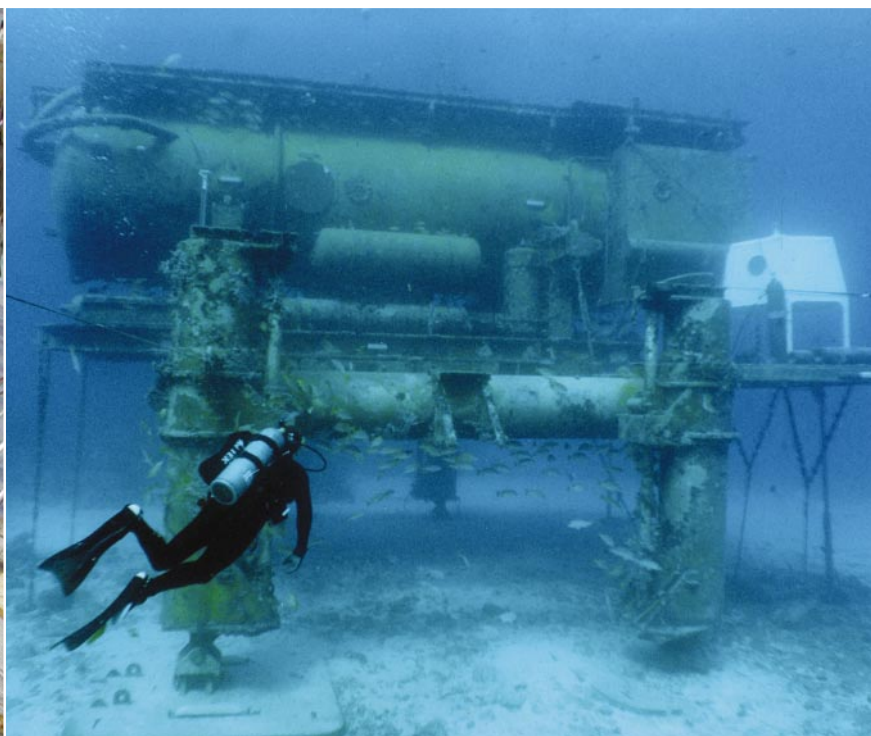
The idea of a network of observation posts for long-term monitoring of the oceans was first realized during the cold war with the US Navy's SOSUS array, a collection of hydrophones draped across the Atlantic and Pacific to listen for Soviet submarines. Its sensitive

ears could pick up low-frequency sounds at less than 1 watt at a range of several hundred kilometres — in the right conditions it could hear the songs of blue whales 3,000 kilometres away. When SOSUS was made available to civilian researchers in the 1990s, it revealed previously unknown seismic rumblings that take place on mid-ocean ridges³ and provided insights into the migration of whales⁴. These results confirmed the scientific value of keeping a constant watch on the oceans.

Scientists now have far more ambitious plans in mind. The boldest is the North-East Pacific Time-series Undersea Networked Experiments (NEPTUNE), the brainchild of John Delaney at the University of Washington in Seattle and Alan Chave at Woods Hole Oceanographic Institution on Cape Cod, Massachusetts. NEPTUNE aims to wire up the entire Juan de Fuca plate — the slab of oceanic crust between the Washington–Oregon coast and the Juan de Fuca Ridge 500 km offshore — with a 3,000-km network of fibre-optic cable. Observatories plugged into 'nodes' on this network will provide a multi-disciplinary playground for ocean scientists, as well as an early-warning earthquake system for the Pacific Northwest.

NEPTUNE is expected to cost a total of \$250 million, but Delaney is adamant about





Science without scuba? Left, researchers deploy a buoy as part of NeMO Net, which will gather data from underwater sensors, providing a convenient alternative to seafloor stations such as Aquarius (above).

the benefits of such a system. "It will fundamentally transform the way that we study the ocean and the Earth, by providing real-time access and data flow to complex processes that are episodic," he says. These episodic events include eruptions on the mid-ocean ridge that seem to release huge plumes of hot water loaded with minerals⁵, which might feed the microbes that live in the water above the vents.

Plans for NEPTUNE have been in the pipeline for over a decade, but it is just now starting to pick up the cash it needs to become a reality. In October 2003, Canada announced Can\$62.4 million (US\$47.5 million) of funding for the project. And hopes are high that a good chunk of the NSF money for ocean observatories might end up in NEPTUNE's pocket.

Meanwhile, both the United States and Canada have funded two small-scale cabled observatories that should serve as test-beds for the project. One is VENUS, the Victoria Experimental Network Under the Sea, which will cover the Straits of Georgia and Juan de Fuca off British Columbia with the first sections of NEPTUNE's cable, scheduled to be laid next year. Keeping to the planetary theme, the other is MARS, the Monterey Accelerated Research System, which plans

to deploy observatories in Monterey Bay, California. Both will allow scientists to keep a better watch on their coastal backyards, while providing an arena to develop NEPTUNE technology.

Stay connected

Delaney does not underestimate the technical challenges faced by regional networks such as NEPTUNE. The foremost is long-term maintenance. NEPTUNE is intended to run for at least two or three decades, with 3,000 km of fibre-optic cable linking 20 or 30 stations, each of which may have secondary extension cords reaching out over tens of kilometres to instrument packages. Keeping this network intact over remote and rugged seafloor terrain is not a trivial operation. In October 1997, the University of Hawaii deployed a seismic observatory on an undersea volcano using 45 km of fibre-optic cable donated by AT&T. Although the observatory piped a wealth of data back to the shore, the cable short-circuited just six months after installation, and has not been brought back online since. A similar fate could lie in store for NEPTUNE cables crossing volcanically active regions of the Juan de Fuca Ridge.

Not all cable-tied projects fail. The Hawaii 2 Observatory (H2O), halfway between Hawaii and California, was built in 1998 using old coaxial AT&T cables and hosts seismometers at a depth of 5,000 metres. Run by the Woods Hole Oceanographic Institution in Massachusetts, it is considered one of the triumphs of deep-sea wired observation, given

the fact that it churns out real-time data from such a remote location. But if researchers do want to get around the cost and potential problems of cables, there is an alternative — beaming data up to surface buoys.

The New Millennium Observatory, or NeMO Net, operated by NOAA, does just that. This 'net' consists of two sensors splayed on Axial Volcano, one of the active underwater volcanoes on the Juan de Fuca Ridge. These sensors beam information about temperature, chemistry and movement of the volcano's flanks by an acoustic signal up to a surface buoy, which provides a satellite link for researchers to receive real-time data and send commands to their instruments. Although satellite communications do not offer the same bandwidth as a fibre-optic connection, and can't feed power to instruments, they can be used in more remote areas without the expense of laying cable.

Meanwhile, on the other side of the world, the European Sea Floor Observatory Network (ESONET) consortium is exploring the possibility of rigging up the Atlantic and Mediterranean coasts. So far ESONET has received E800,000 (US\$983,200) for a feasibility study, which will report in May 2004. Unlike NEPTUNE, ESONET has no tectonic plate to serve as a focus for the project. Instead, the observatories will undertake a range of scientific projects such as assessing the impact of changes in sea ice in the Norwegian Sea on deep-water circulation and monitoring biodiversity in the North Atlantic and seismic activity in the Mediterranean. Ideally,

the project will cover all European waters from the Arctic Ocean to the Black Sea, taking in an area comparable to continental Europe and a host of largely unexplored phenomena from cold-water corals to mud volcanoes.

European researchers already have considerable experience with deep-sea observatories. For example, in August 2001, Monty Priede's team at the University of Aberdeen, UK, deployed an autonomous lander — a mobile observatory that carries its own battery pack and data-storage facilities — at a depth of 4,000 metres for 18 months, to the southwest of Ireland⁶. And European researchers coordinated by the Italian National Institute of Geophysics and Volcanology have used a satellite-buoy observatory to collect oceanographic and geological data from the depths near Sicily for the past seven months.

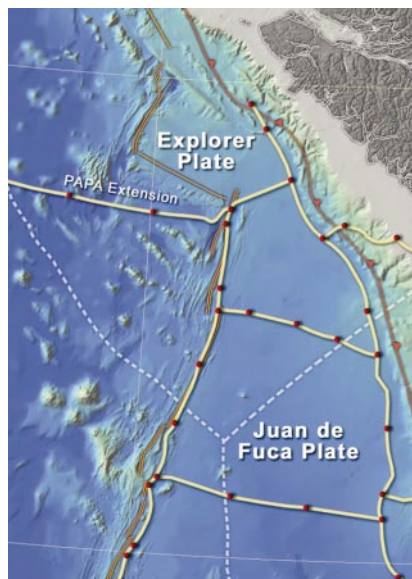
World of experience

ESONET brings together researchers from 14 institutions across the continent with such expertise. Rather than setting their sights on obtaining funding for a full network at the outset, they envisage a gradual growth of capability. "What we perceive is a smooth transition from using autonomous observatories to connecting them up to cables, as opportunities arise," says Priede, who is ESONET's coordinator. "I don't see the the European Commission suddenly coming up with €100 million (US\$123 million) to install ESONET — that's not the way we're going."

Unlike NEPTUNE, ESONET would not be a single integrated network, but a federation of local networks covering different geographic regions. "We're not pretending that there will be a cable connected all the way from Norway to Sicily," says Priede. "But on a 20-year timescale, I think we'll have a very significant monitoring capability around Europe."

To make this happen, ESONET members are exploring the possibility of sharing infrastructure with other partners, including astronomers. Researchers now working on a project to build a neutrino telescope 2,400 metres below the surface of the Mediterranean Sea⁷, for example, have already set up a cable connection to the shore to beam back their data. And that could be shared with seafloor observatories in the area.

The third big player in ocean observatories is Japan, which has traditionally been interested in picking up signals from seismic instruments on the sea floor to give an early warning of earthquakes and tsunamis, the giant waves they produce. Eight cabled observatories are already in place around the islands, mostly for earthquake observation. And researchers are planning the more ambitious ARENA (Advanced Real-time Earth Monitoring Network in the Area). Headed by Yuichi Shirasaki at the University of Tokyo, the ARENA project aims to provide a network of cabled observatories spanning the plate boundary along



Underwater empires: the NEPTUNE and ESONET projects will position sensors around the US (left) and European (top right) coasts, respectively, studying everything from earthquakes to coral (above).

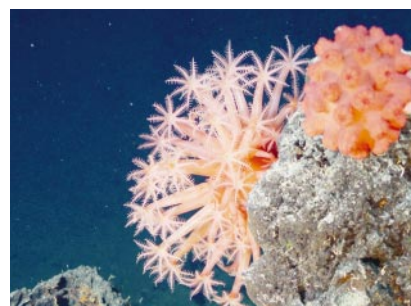
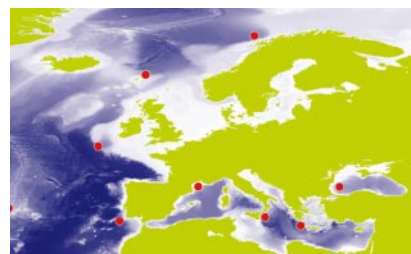
the Japan Trench. As with NEPTUNE and ESONET, researchers anticipate that the observatories will be multidisciplinary, allowing researchers to collect a suite of seismic, oceanographic and biological information.

While NEPTUNE, ESONET and ARENA are intended to probe the deeps, researchers in shallower waters have been enjoying the benefits of seafloor observatories for some time. Rutgers University, for example, runs LEO-15, a long-term ecosystem observatory lying 16 km off the New Jersey coast. Its two nodes, which were designed by the Ocean Systems Laboratory at Woods Hole, Massachusetts, host sensors that monitor temperature and currents, and take video images to stream back to the lab.

Fishing for clues

LEO-15 has proved eminently useful not only to scientists but also to fishermen along the East Coast. In 1976, the shellfish industry was devastated by a mysterious shortage of oxygen in the water. Pollution was the first suspect, but the culprit turned out to be a strong upwelling which brought an influx of nutrient-rich cold water towards the coast from deeper off-shore. This caused blooms of phytoplankton that later died and decayed, consuming much of the oxygen in the water. LEO-15 now gives fair warning of incoming upwelling waters, allowing the fishermen to plan for difficult times ahead.

One node of LEO-15 acts as a service station for the autonomous underwater vehicles that prowl around the observatory. Woods Hole's remote environmental sensing unit (REMUS) can dock with the node to download its data and recharge its batteries. Delaney expects a similar squadron of vehicles to scurry about the Juan de Fuca Plate like bees, visiting the nodes of the NEPTUNE network and col-



lecting data from the gaps between them. The robotic craft could also be used to respond to episodic events detected by the network, for example moving in to sample and survey the epicentres of seismic activity.

Rutgers University also plans to hook LEO-15 up with similar projects, such as the Martha's Vineyard Coastal Observatory, to create a monitoring network covering most of the eastern seaboard. The project, nicknamed Cable Tie, plans to use retired transatlantic telephone cable to link the observatories. By combining underwater observatory data with satellite images, observations from radars that monitor ocean-surface currents, measurements from oceanographic buoys and from vehicle missions, the network hopes to work out the carbon cycle of the continental shelf.

While the dream of life beneath the waves remains elusive, marine scientists are spending more and more time immersed in their medium through such 'telepresence'. Those with plans to extend this capability further are adamant about its benefits. "We've come a long way with the standard approaches of ships and satellites, but we've reached a stage where we cannot go much farther without this additional capability," Delaney argues. "That's not to say that this will replace ships or satellites by any means, but it's the next tool that we desperately need in order to move on."

Jon Copley is a freelance writer based in Southampton, UK.

1. Gruber, N., Keeling, C. D. & Bates, N. R. *Science* **298**, 2374–2378 (2002).
2. Dore, J. E., Lukas, R., Sadler, D. W. & Karl, D. M. *Nature* **424**, 754–757 (2003).
3. Fox, C. G. et al. *Geophys. Res. Lett.* **22**, 131–134 (1995).
4. Charif, R. A., Clapham, P. J. & Clark, C. W. *Marine Mammal Science* **17**, 751–768 (2001).
5. Cowen, J. P. & Baker, E. T. *Deep-Sea Res. II* **45**, 2503–2511 (1998).
6. Bailey, D. M., Jamieson, A. J., Bagley, P. M., Collins, M. A. & Priede, I. G. *Deep-Sea Res. I* **49**, 1519–1529 (2002).
7. Ball, P. *Nature* **419**, 12–14 (2002).

Tough lessons for survival in hard academic times

Should young scientists still follow their hearts or do they have to follow the money?

Sir — Steven Weinberg's Concepts essay "Four golden lessons" (*Nature* **426**, 389; 2003) is full of idealism, based on his experience, garnered "about a hundred years ago". Sadly, the research and economic worlds have changed dramatically during the past quarter-century. I suggest that Weinberg's rules should be revised for modern would-be postgrads.

One: look at the career structure in scientific research — it is virtually non-existent. Research careers are usually tied to teaching, so if you want to forge a future in research then you will need to secure academic tenure. If you are still dependent for your salary on 'soft' money — research grants — by the age of 35, you will then be told by (much older) tenured colleagues that you are "too old" for research and that you should look for another career. So see your early steps into the research world as leading towards a completely different career. Banking, finance or teaching are common end-points. Academic administration may provide a means for

revenge against those professors who misled you about your future.

Two: take note of which areas of research in your chosen discipline have the oldest entrenched academics, and head for those. Many were filled in the 1970s by baby-boomers who are now approaching retirement, so you may be well-positioned for one of their jobs.

Three: look at the best jobs outside academia. For example, a well-known scientific journal advertised a research position last year in a British astrophysics department. Conditions included a poor salary of less than £20,000 (US\$35,000) a year, and limited tenure for 12 months "with the possibility of renewal for a further 12 months". The position required a maths/physics postgraduate with extensive experience in database management. On the following page was an advertisement from a London merchant bank. This asked for identical qualifications, but promised a starting salary of £48,000 "with rapid advancement" and a well-structured career

pathway. So there are good career opportunities for postgrads. They just don't happen to be in academia.

Four: look at the new fields emerging for employment in big, profitable industries. For example, the pharmaceutical industry employs many graduates, in lab research, database and analysis, clinical trials and marketing. Annual reports will reveal what fields companies are moving into, and what they are dropping. 'Pharmacogenomics' and 'proteomics' are examples of trendy new fields that are attracting large budgets, whereas animal testing is gradually being wound down in favour of *in vitro* cell modelling and large-scale, mathematically based analysis such as cladistics.

Choose your research path according to hard-headed economics, and forget the good old days when students went into research because it was fun. You know that things are different now.

John A. Duley

Mater Misericordiae Hospital, Raymond Terrace, South Brisbane, Queensland 4101, Australia

White House cost-cutting undermines productivity

Sir — Having read the News story "Democrats condemn government 'meddling' with NIH" (*Nature* **425**, 888; 2003), I was overcome by disgust for the blatant disregard the White House shows towards education and research. I find it difficult to understand the logic behind privatizing a historically government-run agency. It is true that you could argue that the National Institutes of Health should be run more efficiently, but you could argue that the White House should be run more efficiently too. Education and scientific advancement are the backbone of any great nation. Applying pressure and weakening morale in these areas will only lead to declined productivity, and a weaker United States of America.

A News in Brief story in the same issue, "Foreign students to foot the bill for US security scheme" (*Nature* **425**, 892; 2003) describes the White House plan for 'taxing' foreign-student visas as a way to raise capital for homeland security. I work in a research institution that employs a large number of foreign research associates whose work advances medicine and science, often with the goal of saving lives. I have seen accomplished foreign scientists encounter situations in which immigration has made employment in the United States such a headache that they chose to take

their skills to a different country.

At what point will the decision-makers in Washington realize that their 'initiatives' cause more grief than good?

Matthew C. Salanga

Division of Neuroscience, Children's Hospital, 320 Longwood Avenue, Boston, Massachusetts 02115, USA

Joint efforts needed to forecast space weather

Sir — In response to your News story "Bleak forecast for space weather" (*Nature* **425**, 649; 2003) about US funding and organizational issues, we would like to suggest possible European and global collaboration in this field.

Scientists in the United States and Europe now believe that computer-based forecasting of space weather is possible in real time, from hours to possibly a few days ahead. Such forecasts would provide estimates of the effects of solar flares and solar particle fluxes on Earth's magnetic field. Computer and data-handling systems would be similar to those used for numerical weather prediction but would instead use the equations of magneto-hydrodynamics and plasma physics and rely on real-time satellite (for example, SOHO, CLUSTER, GEOS, ACE, RHESSI, TRACE, WIND, POLAR and

GEOTAIL) and ground-based observations.

Regional forecasting centres would be needed to operate these systems. We proposed one such centre — the European Centre for Medium Range Weather Forecasting — at a recent Royal Society meeting (J. C. R. Hunt and A. J. Coates *Phil. Trans R. Soc. A* **361**, 205–218; 2003), and another was proposed as a possible role for the National Center for Atmospheric Research in the United States, following a review of the centre chaired by James Baker in November 2001.

Recent events such as the unusual solar activity seen during the past few months (*Nature* **426**, 112; 2003) have shown that quantitative predictions are needed for how solar activity affects systems on Earth, including air-traffic control and electrical power grids. Just as plans are developing to organize global warnings of near-Earth objects in space, surely there are equally good reasons to establish a global cooperative network for forecasting space weather, involving major centres in North America, Europe and Asia?

Julian Hunt, Andrew Coates

Department of Space and Climate Physics, University College London, Gower Street, London WC1E 6BT, UK

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

The hidden cost of health care

Is biomedical research justified on moral and ethical grounds?

What Price Better Health? Hazards of the Research Imperative

by Daniel Callahan

University of California Press: 2003. 335 pp.
\$29.99, £19.95

Arthur Caplan

This book is intended for a wide general audience, but if you are a biomedical scientist, you would be well advised to read it. For the philosopher Daniel Callahan presents what might just be the strongest case for putting you out of business.

Callahan founded the Hastings Center, a bioethics research centre in New York state, and was its director for three decades. For much of that time he has wondered about the value of biomedical research. *What Price Better Health?* constitutes the crescendo of his dissatisfaction with the biomedical enterprise.

Biomedical research has set itself up inside an unassailable moral fortress, argues Callahan. Its main moral weapon is what he dubs the 'research imperative' — the idea that research must proceed regardless of the cost or risk. No downside can trump the research imperative, because a cure for all that ails us, and possibly a potion to stave off death itself, could be contained within the next grant proposal, embedded in the next biotech start-up, or just around the corner of the next widely touted new idea (be it fetal-tissue research, gene therapy, embryo research, xenografting, artificial organs, stem-cell research or genomics, for example). What hope is there for those who want to challenge the way that biomedical research is done, when to do so they must stand in the way of the research imperative, and thus in the path of progress?

In this book, Callahan attempts to slay the dragon of the research imperative with two main arguments. First, he says that the research imperative has completely corrupted biomedical science. Second, he argues that the research imperative is no imperative at all.

In chapter after chapter he argues that the price of biomedical science's drive to find cures for everything is corruption. The integrity of science in terms of honesty, openness and cooperation has completely frayed. The rights of humans in research are constantly being compromised or ignored. Conflicts of interest abound. Advocacy groups, regulatory bodies and government commissions simply roll over and play dead in the face of the promises of the research juggernaut. And lording over the whole unseemly mess, he says, is the bloated, greedy figure of the pharmaceutical industry, which earns huge profits while addicting us to drugs



Dash for cash? But critics of biomedical research are seen as standing in the way of medical advances.

that don't work or that we don't really need.

Not only is the price exacted by obedience to the research imperative far too high, but the imperative itself is a mirage. Callahan argues that there is no moral obligation to undertake biomedical research. First, he says that we have no duty to try and improve things for our descendants. Then he adds that although health is an important goal, it is far from being the only good that we have an obligation to produce.

So does he make his case? Should we try to shut down the US National Institutes of Health (NIH), INSERM (France's national biomedical agency), Britain's Medical Research Council and the pharmaceutical industry before they destroy us? Should we abandon the quest for eternal vigorous life and come to stoical grips with our finitude and frailty? Callahan displays a real mastery of policy and history but, despite his efforts, the research imperative does not yield to his rhetorical sword.

Is it really the case that the research imperative has a Svengali-like effect on all who hear its dulcet tones? Hardly. Even in the United States, whose government funding constitutes a good chunk of the world's share of biomedical research and development (R&D), the total expenditure on biomedical research, even though it has grown over the past decade, is barely a detectable blip in the US economy.

Last year the US government spent more than \$2 trillion in total, of which \$107 billion went on R&D. More than half of that amount went on military and defence-related research.

The NIH budget was roughly \$27 billion but a sizeable chunk of that was targeted at research on topics relating to bioterrorism and national security. Contrary to Callahan's view that Americans are digging deep into their wallets in egomaniacal hope of eternal life, their actual expenditure seems to show that they are well aware that there are lots of other things to spend money on, and they do.

What about the duty to make things better for those who will follow us? It is true that moral theories do not posit an obligation to improve the lot of those who have not yet been born. But we do have an obligation to try and make things better for our own children, and biomedical research is a crucial element of how that obligation should be discharged.

More to the point, we may also have a duty to support biomedical research because we choose to benefit from the fruits of past research. If I use a hearing aid, take insulin for my diabetes, walk using an artificial hip, wear a pacemaker, see through eyes shaped by laser surgery, and vaccinate myself against diseases, then I am benefiting from previous investments in biomedical research. It would seem that, unless I want to act as a 'free rider', I owe a debt to the investments of my forebears that must be discharged by continuing to support biomedical research. Fair play demands that I pay for what I benefit from.

If we do have a duty to do research, that brings us to the other part of Callahan's argument: has the biomedical research imperative crushed key values and core moral principles? Here Callahan's argument

T. & D. A. MCCARTHY/CORBIS

has some traction. Things are not as they should be in terms of internal values in biomedicine. Repairs are certainly in order.

Surprisingly, though, Callahan offers no comment on one of the most potent threats to scientific integrity and values — the power of government-sponsored military and anti-terror research to undermine the integrity of science. This is not a new problem but recent events have raised its profile. In the United States and Europe, some of the greatest threats to key scientific values come from the desire of government to keep secret the work that it funds in the name of national security.

But having said that, do proponents of biomedical research really wield the research-imperative weapon in the way that Callahan maintains? Most biomedical researchers are keenly alert to the obligations to treat human and animal subjects respectfully and with dignity. They understand the tensions imposed by private funding on the ethos of their work. And they are open to listening to and taking seriously the objections of those who fret about where biomedical technology might take us. And so they should. Despite Callahan's hyperbole about the power of the biomedical research juggernaut, critics have scored some victories. The genetic modification of plants and animals is moving much more slowly than proponents would like; stem-cell, embryo and cloning research are being subjected to close scrutiny; and efforts to advance xenografting and the creation of artificial hearts have come more or less to a grinding halt for a variety of ethical and social reasons.

Callahan has written an important book. The research imperative may not be quite as invulnerable as he thinks, but it is certainly imperative that the case he makes against it be given the close and thoughtful attention that his book provokes.

Arthur Caplan is in the Department of Medical Ethics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104-3308, USA.

Characters from the dawn of chemistry

The Last Sorcerers: The Path from Alchemy to the Periodic Table

by Richard Morris

Joseph Henry Press: 2003. 296 pp. \$24.99

John Emsley

Who was the last true alchemist? Probably Johann Frederick Böttger (1682–1719), who started out looking for the Philosopher's Stone and ended up finding a way to make porcelain. Who was the first real chemist? Probably Robert Boyle (1627–91). He also began as an alchemist but became a chemist

when he turned his attention to the newly discovered phosphorus in the 1670s. This he investigated in a systematic way, and he published his findings not in the arcane language of the alchemists, but in plain English.

This approach showed how far from alchemy Boyle had come, although he still believed that transmutation — turning one metal into another — might be achieved. For this reason, he fell prey to a scam wrought by a Frenchman, George Pierre des Clozets, who wrote to Boyle telling him that he could join a secret society of alchemists — for a fee. The upshot was that des Clozets milked Boyle of large sums of money. It is the inclusion of this kind of anecdote that makes Morris's book such a fascinating read.

The Last Sorcerers is a collection of short biographies of key individuals who span the years that saw the end of alchemy and the emergence of chemistry. It starts with an excellent account of Paracelsus and ends with one about Niels Bohr. Along the way we meet such chemistry greats as Antoine-Lurent Lavoisier, Henry Cavendish, Joseph Priestley, Jöns Jacob Berzelius and Dmitry Mendeleev. In every case, Morris writes with a nice blend of science and human interest. I kept hoping that Morris might find a common thread of personality to unite his characters — what drove such a diverse bunch of men to study chemistry? Nothing illustrates the contrast more than the discoverers of hydrogen and oxygen. The former was the richest man in England; the latter was hounded out of the country for his radical left-wing views.

Cavendish was so wealthy that the Bank of England held his money in a special account. But he was so unworldly that when the bank sent a representative to suggest that he invest some of the £90,000 (equivalent to about £20 million, or US\$35 million, today) that had accumulated in his account, he sent the man away saying he didn't want to be "plagued" about it, so there it sat growing ever larger. All Cavendish needed was enough money to enable him to carry out his experiments in his private laboratory in Clapham, and indeed he so little understood money that he gave the man whom he had employed to catalogue his library a cheque for £10,000. Cavendish was a recluse and was terrified of women, yet he performed some remarkable experiments that changed the course of chemistry, most notably making, collecting and studying hydrogen gas.

Priestley did the same for oxygen, but he was a non-conformist preacher who was married and relatively poor. He wrote inflammatory pamphlets in support of the French and American revolutions, and was attacked not only by the press, but also by rioters in Birmingham, who burned down his house and laboratory.

Not all chemists were so badly treated;

some were even admired and loved. When John Dalton died, aged 74, in Manchester in 1844, the city fathers had his body taken to the Town Hall, where some 40,000 citizens filed past it to pay their last respects. The following day his funeral procession was a mile long, with 100 carriages and tens of thousands of ordinary people following on foot.

The Last Sorcerers is well-written popular science, and as such deserves to be widely read. That it deals with chemistry's somewhat shady origins adds to its attraction. The fact that it also reveals the human side of some famous chemists adds even more to one's enjoyment.

John Emsley is an author of popular chemistry books, his latest being *Nature's Building Blocks* (Oxford University Press).

The bits that make up the Universe

Information: The New Language of Science

by Hans Christian von Baeyer

Weidenfeld & Nicolson: 2003. 258 pp. £16.99

Michael A. Nielsen

What is the Universe made of? A growing number of scientists suspect that information plays a fundamental role in answering this question. Some even go as far as to suggest that information-based concepts may eventually fuse with or replace traditional notions such as particles, fields and forces. The Universe may literally be made of information, they say, an idea neatly encapsulated in physicist John Wheeler's slogan: "It from bit". Others rather less boldly suggest that taking a point of view based on information theory may yield insights into existing theories such as statistical mechanics and quantum mechanics.

These are speculative ideas, still in the early days of development. Their most encouraging success is perhaps the resolution of the 'Maxwell's demon' paradox, a century-old riddle in the foundations of statistical mechanics. In James Clerk Maxwell's paradoxical thought experiment, a demon of extraordinary dexterity and visual acuity partitions an initially homogeneous gas into two parts, one part containing slow-moving molecules and the other part faster-moving ones. In the thought experiment, the gas is initially spread evenly through a two-chamber container with a connecting trapdoor that can be opened and closed by the demon. By carefully observing the velocity of molecules approaching the trapdoor, and opening or closing it as appropriate, the demon sorts the molecules so that fast molecules enter one chamber and slow ones end up in the other.

Virtual art

Art that draws you in

Computer-generated virtual reality is a new form of art only in terms of its medium, according to art historian Oliver Grau of the Humboldt University in Berlin, Germany. In his highly original book *Virtual Art: From Illusion to Immersion* (MIT Press, \$45), Grau argues that artists have been using a variety of techniques to immerse observers within their works for millennia.

The earliest trick simply involved physically surrounding the viewer with images. The frescos covering all four walls in a room in the Villa dei Misteri at Pompei, Italy, created in about 60 BC, draw observers into a 360° depiction of the preparations for a cult ritual. Those adorning a room in Livia's Villa at Prima Porta, dating from about 20 BC, create a 360° illusion of a garden. Post-antiquity, the frescos in the *Chambre du Cerf* (Chamber of the Stag) at Pope Clement VI's palace in Avignon, France, from 1343, place the observer at the centre of a hunting scene.

Mathematical perspective was the visual trick of the Renaissance. It is epitomized by the sixteenth-century *Salle delle prospettive* (Chamber of Perspectives) at the Villa Farnesina in Rome, the walls of which depict a columned hall. Between the pillars of the portico appears a view looking out onto the city and to the hills beyond.

Peepshow boxes — the forerunners of the stereoscope and the head-mounted display — first appeared in the eighteenth century, and in 1787 Robert Barker patented his hugely successful process for producing panoramic views on circular canvasses in correct



Virtually there: computers allow viewers to immerse themselves in Charlotte Davies' *Osmose*.

perspective for observers standing at the centre. The techniques of cinematography and computing simply extended the opportunities for artists to immerse observers in an illusory world, according to Grau.

In the 1990s, the developers of virtual reality began to hire artists to assist them. Among these was Charlotte Davies, whose *Osmose*, shown here, is a total immersion in the fabric of the living Earth — rocks, roots, trees and leaves. It is a product of her relationship with the Canadian software company Softimage.

But digital artworks are vulnerable to

extinction as the operating systems on which they are based become redundant. Grau is cooperating on an international, interdisciplinary level with art academies and research laboratories to document two decades of computer-based art, much of which already cannot be shown. He has, for example, built a database of virtual art, a cataloguing project that is part technological and part art-historical, and is intended to support preservation efforts. *Osmose* is one of hundreds of works that will become publicly accessible through this database in the coming months. **Alison Abbott**

The resulting system is more ordered than the original homogeneous gas, and so has lower entropy. Furthermore, by making the trapdoor sufficiently lightweight, the demon can operate it by expending an arbitrarily small amount of energy. Thus, a naive analysis suggests that Maxwell's demon reduces the total entropy of the system, violating the second law of thermodynamics.

This paradox was resolved in 1982, when physicist Charles Bennett, building on earlier work by others, notably Rolf Landauer, showed that this analysis fails to take into account an entropy cost associated with the information acquired by the demon when it observes the velocities of molecules approaching the trapdoor. The cost is an entropic price paid when the demon erases its record of these observations. Remarkably, when this cost is taken into account, the violation of the second law is found to be illusory and the paradox is resolved.

A more recent example of information taking a surprising central role in fundamental science is a bold idea known as the holographic principle. Roughly speaking, this states that the correct way to describe a region of space-time is not through a

description of fields and forces in the bulk of the volume, as is conventionally done, but through a theory whose elements are defined on the surface of the region. The motivation for the principle comes in part from results about the thermodynamics of black holes suggesting that the information content of a black hole is proportional to its surface area, not its volume. Some researchers hope that the holographic principle will help lead to a quantum theory of gravity, much as Einstein's principle of equivalence helped to motivate the general theory of relativity.

These and other examples illustrate the intellectual ferment associated with the role of information in fundamental science. It is against this backdrop that Hans Christian von Baeyer's elegant popular book is set.

The book's most appealing feature is its focus on big questions. What is information? What role does information play in fundamental physics? Where else in science does information play a critical role? And what common themes link these areas? Von Baeyer approaches these questions from many angles, giving us a flavour of some of the most interesting answers currently being offered.

There is a nice balance between accepted science and speculative ideas. For example, the standard theory of information, proposed by Claude Shannon in the 1940s, is introduced early in the book. However, von Baeyer admits that Shannon's theory has some shortcomings, and provides a flavour of several other approaches to developing information theories, notably quantum information theory.

Von Baeyer discusses many fascinating topics in a tour that is broad but not deep, taking in genetics, bioinformatics, quantum computation, the foundations of quantum mechanics, and black-hole entropy. He faces, and on the whole overcomes reasonably well, the difficulty faced by popular science writers of needing to simplify without misleading. However, I did notice several unfortunate minor errors of fact.

In summary, von Baeyer has provided an accessible and engaging overview of the emerging role of information as a fundamental building block in science. ■

Michael A. Nielsen is in the School of Information Technology and Electrical Engineering and the School of Physical Sciences, University of Queensland, Queensland 4072, Australia.

Water, water, everywhere?

Philip Ball

On Earth, no living organism can function without water. It is, in the words of Albert Szent-Györgyi, the matrix of life. But is it reasonable to assume that this maxim holds on other worlds too?

Is life possible without water? NASA has stated explicitly that its strategy in searching for extraterrestrial life is to “follow the water”. But is the space agency thereby overlooking other potentially fertile environments? A recent meeting* of physicists, chemists, biochemists and microbiologists grappled with the question of whether water-free life is feasible. No one can give a definitive negative answer, and neither can we expect the issue to be resolved by a show of hands. Rather, the task has to be that of reducing the basic question to smaller, tractable ones, in the hope that a framework might emerge for moving the discussion beyond mere speculation.

The naive response might be to suppose that the question is absurdly terracentric. If one allows — and it seems a reasonable, though not invulnerable, starting point — that a liquid of some kind is required simply for efficient mass transport in living systems, the cosmos could provide plenty of alternatives: ammonia, sulphuric acid, liquid carbon dioxide, even the putative hydrocarbon lakes of Saturn's moon Titan.

But there is much more to water than that. It has long been recognized as a profoundly anomalous liquid, with properties that set it apart from all others. High heat capacity, expansion on freezing, maximum density at 4 °C, high dielectric constant — all of these so-called anomalies, and others, seem critical to its biological role. They are in fact relatively easy to rationalize on the grounds of water's hydrogen-bonded structure, which joins the H₂O molecules into a fluctuating, three-dimensional network (J. Finney, University College London). Unlike ‘simple’ liquids, water's molecular structure is dominated not by the hard core repulsions between molecules but by the directional, attractive interactions of hydrogen bonds.

Is this unusual character an essential, or just an incidental, factor in water's life-giving agency? The mission of the meeting was to identify the molecular aspects of water's role in life on Earth, and then to ask whether there was any reason to regard these properties as generic or optional. And if the former, could they be reproduced by any other liquid?

The apparent ‘specialness’ of water was pointed out in 1913 by the American biochemist Lawrence Henderson, who argued that the Universe seems remarkably ‘fit’ to



The meaning of life? Water molecules are involved in the biology of all life-forms on Earth. That might not be true on other planets.

foster life — a precursor to the anthropic principle. But there is an inherent danger of circularity here: because life is adaptive, who is to say that it has not simply found ways to exploit what water has to offer? For example, some proteins make use of the fast proton conduction that takes place in water, a consequence of bond-flipping along chains of hydrogen-bonded molecules. (The details are, however, more complicated than implied by the classical Grotthuss mechanism; see N. Agmon *Chem. Phys. Lett.* **244**, 456–462; 1995.) Some proteins use one-dimensional chains of water molecules to carry protons rapidly to active sites in their interior. The hydrogen-bonded network makes water particularly well suited to providing such ‘proton wires’. But if this trick were not available, is there any reason to suppose that life would be stymied?

One can postulate that life of any sort will require enzyme-like selectivity of molecular interactions for transmitting chemical information. Water does seem to play many subtle parts in enzyme function, but is it really irreplaceable? Solvation shells can be seen to be active components in protein function (J. Smith, Univ. Heidelberg; M. Nakasako, Keio Univ.; P. Rand, Brock Univ.). But it is not obvious that other small-molecule solvents could not substitute, in principle.

In fact, studies of enzymes in low-water environments give rather conflicting messages about the extent to which water is needed. The bacteriorhodopsin protein, embedded in its natural ‘purple membrane’, seems to switch on only when there is at least a monolayer of water hydrating the exposed protein surface (G. Zaccai, Inst. Biologie Structurale, Grenoble), whereas some enzymes work in the gas phase without any hydration layer at all (R. Daniel, Univ. Waikato). Although experience with non-aqueous solvents has given some researchers confidence that enzymes will ultimately be found that work efficiently entirely without water (D. Clark, Univ. California, Berkeley), part of the difficulty here is that ‘water-free’ means different things to different people. At present, all functional enzymes seem to retain ‘internal’ water bound strongly inside their protein structure (at concentrations as low as one to ten H₂O molecules per mole of protein) — can that, too, be removed? Whereas hydration-shell water seems mainly to promote flexibility (and may be fully replaceable by other solvents), internal water seems to preserve the protein's conformation. It can, perhaps, be ‘designed out’ of the system, but not easily.

Yet how can molecules that have evolved in water tell us about what is possible without

*The Molecular Basis of Life: Is Life Possible Without Water? The Royal Society, London, UK, 3–4 December 2003.

it? They might at least help to narrow the question: if a wholly water-free enzyme were found, we could feel confident that at least this molecular aspect of life need not rely on water's uniqueness. Similar arguments apply to protein folding: that is, making the catalysts in the first place. Investigating water's role here reveals many subtleties. For example, a good solvent doesn't actually promote stability of the native fold — the conformation that a protein naturally assumes. Rather, it finds a remarkably delicate balance between strong, conflicting forces so as to promote only marginal stability (J. Goodfellow, BBSRC). If there is too much stability the structure 'freezes' and becomes inactive. The alternative protein conformations revealed in amyloid diseases may simply be the inevitable price that we pay for this.

There seems to be no simple molecule

that can mimic all of the useful biological functions of water. One school of thought asserts that it is therefore futile to look for replacements for any one, or even simultaneously for several, of its 'virtues': the biological importance of water lies in their synchronous operation in a single molecular system. But what we really need is a way of asking which, if any, of those functions is generic to life. Is there, for example, a temperature limit that rules out other tetrahedral liquids such as silica, because of the complications introduced by molecular excited states at high temperatures? At low temperatures, would slower diffusion rates prevent effective exploitation of thermodynamic equilibria? In other words, is there a habitable zone not just in physical space but in chemical and thermodynamic space too? ■

Philip Ball is a consultant editor for Nature.

Vision

The need for speed

Kendall J. Blumer

Neurons in the retina turn on and off rapidly in response to light. With the discovery of mutations in human genes that mediate this quick turn-off, we have the first picture of its importance in visual perception.

Imagine walking out of a dark theatre into a bright and sunny Sunday afternoon. You are momentarily blinded, but your eyes rapidly adjust to the change and you continue on your way. For some people with a rare visual defect, however, this momentary blindness can last for up to ten seconds. A similar, but potentially more dangerous, prolonged blindness occurs when these individuals drive from daylight into a darkened tunnel. Moreover, people with this problem also suffer from difficulties in seeing certain moving objects (such as balls thrown during

a sporting event). On page 75 of this issue, Nishiguchi *et al.*¹ describe a genetic cause of this condition. In so doing, they reveal that visual perception requires rapid deactivation of the light-stimulated responses shown by neurons in the eye.

Light streaming into the eye is detected by specialized neurons (photoreceptors) in the retina. In response to light, a coordinated series of molecular events — the so-called phototransduction cascade — is triggered in these cells² (Fig. 1). Photons excite pigment-containing proteins called rhodopsins, which

then switch on the protein transducin by loading it with the small molecule guanosine triphosphate (GTP). When bound to GTP, transducin turns on a phosphodiesterase, an enzyme that breaks down cyclic guanosine monophosphate (cGMP — another small molecule). High concentrations of cGMP open specialized ion channels in the outer cell membrane. Thus, by reducing the concentration of cGMP, light changes the flow of ions across the membrane of photoreceptive neurons, producing an electrical signal that is necessary for communicating with the brain.

Once this light-activated switch is on, how do cells turn it off? One mechanism is to limit the amount of time that GTP-bound transducin can keep the phosphodiesterase enzyme active. Transducin can accomplish this task itself by converting — hydrolysing — its bound GTP molecule into guanosine diphosphate, GDP. (This conversion from GTP to GDP is a commonly used molecular 'switch' in a variety of cellular signalling pathways.) Because transducin bound to GDP has a low affinity for phosphodiesterase, it releases the enzyme in an inactive form, allowing cGMP levels to rise again and return the flow of ions across the cell membrane to the 'dark' state. In this molecular cascade, then, the conversion of GTP to GDP by transducin is the rate-limiting step that defines the amount of time for which a photoreceptor responds to a light pulse.

But this presents a problem. Photoreceptor cells can turn off in less than a second in response to a brief flash of light². In contrast, the hydrolysis of GTP by transducin requires tens of seconds to complete, making it difficult to understand how such a mechanism could account for the rapid turn-off of photoreceptor cells. To get around this problem, photoreceptor cells possess a protein called regulator of G-protein signalling 9 (RGS9) that accelerates transducin's ability to hydrolyse GTP³. Indeed, mice that lack the RGS9 gene exhibit slow photoreceptor deactivation⁴.

Building on these studies of mice, Nishiguchi *et al.*¹ now show that disruption of this accelerator mechanism is the likely cause of a 'slow photoresponse recovery' condition previously described⁵ in several humans. The authors started by analysing the DNA of five unrelated people with the condition, and found that four of them had mutations in both copies of their RGS9 gene, producing a protein that is a poor accelerator of GTP hydrolysis. In the fifth patient, the RGS9 gene was normal. Instead, this person had a mutation that inactivates the R9AP gene, which encodes a retinal protein that anchors RGS9 to membranes⁶.

The identification of these mutations also provided an opportunity to study their effects on visual perception — something that, for obvious reasons, could not be studied in mice. The visual abnormalities of these patients

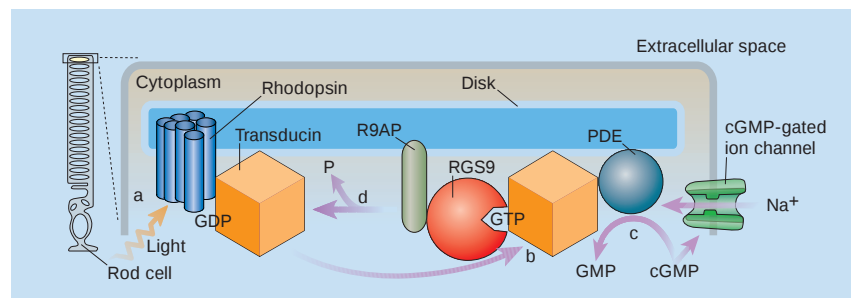


Figure 1 Phototransduction in photoreceptor cells. In the rod class of photoreceptors, the pigment-containing protein rhodopsin absorbs light (a) and activates transducin (b) by causing it to release GDP and bind GTP. GTP-bound transducin binds to and activates a phosphodiesterase (PDE), which converts cGMP to GMP (c). The concentration of cGMP decreases below what is required to open cGMP-gated ion channels, reducing the flow of ions across the cellular membrane. RGS9 bound to R9AP turns off the light-induced response by accelerating the rate of GTP hydrolysis by transducin, releasing phosphate, P (d). Other proteins that regulate the phototransduction cascade have been omitted for clarity. Nishiguchi *et al.*¹ have identified several people with mutations in RGS9 or R9AP. These patients show slow photoreceptor deactivation and have difficulty in adjusting to changes in light levels, as well as in seeing low-contrast, moving objects.

were striking: all reported difficulties in adjusting to changes in light intensity (such as when walking out of that dark theatre). Measurements of the retinal activity in these individuals confirmed that adaptation to changes in dim or bright light, or recovery from a bright flash of light, was much slower than normal. Some patients reported that they could not participate in sport because they could not see a moving ball. In agreement with this complaint, the patient who had a mutation in R9AP also had greatly impaired visual acuity (20/200) for moving, low-contrast objects, although visual acuity was nearly normal in standard tests using high-contrast objects. The findings indicate that adaptation to changing light conditions, visual acuity and contrast detection require that photoreceptors be able to deactivate rapidly. Contrast perception is likely to be impaired because the retinal cells that relay signals to the brain respond most strongly to rapid changes in input from photoreceptor cells. Given these findings, Nishiguchi *et al.*¹ propose calling this condition bradyopsia ('slow vision').

RGS9 is one of nearly 30 such RGS proteins, which regulate signalling by hundreds of receptors coupled to transducin-like G proteins in cell networks of the nervous, cardiovascular, sensory and immune systems⁷. In addition to its function in the retina, RGS9 is highly expressed in the rodent brain and is involved in controlling the rewarding effects

of dopamine and the pain-killing effects of morphine^{8,9}. Human RGS4 is also expressed in the brain and has been linked to schizophrenia¹⁰, whereas rodent RGS2 is widely expressed and regulates immune function, aggressive behaviour and blood pressure^{11,12}. Nishiguchi and colleagues' study¹, however, represents the first description of a human condition that is linked to reduced RGS function. It seems likely that more examples of human disorders that are caused by impaired RGS function will be discovered as our understanding of these signalling molecules increases. ■

Kendall J. Blumer is in the Department of Cell Biology and Physiology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA.

e-mail: kblumer@cellbio.wustl.edu

1. Nishiguchi, K. M. *et al.* *Nature* **427**, 75–78 (2004).
2. Arshavsky, V. Y., Lamb, T. D. & Pugh, E. N. Jr *Annu. Rev. Physiol.* **64**, 153–187 (2002).
3. He, W., Cowan, C. W. & Wensel, T. G. *Neuron* **20**, 95–102 (1998).
4. Chen, C. K. *et al.* *Nature* **403**, 557–560 (2000).
5. Kooijman, A. C., Houtman, A., Damhof, A. & van Engelen, J. P. *Doc. Ophthalmol.* **78**, 245–254 (1991).
6. Hu, G. & Wensel, T. G. *Proc. Natl Acad. Sci. USA* **99**, 9755–9760 (2002).
7. Hollinger, S. & Hepler, J. R. *Pharmacol. Rev.* **54**, 527–559 (2002).
8. Rahman, Z. *et al.* *Neuron* **38**, 941–952 (2003).
9. Zachariou, V. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 13656–13661 (2003).
10. Mirnics, K., Middleton, F. A., Stanwood, G. D., Lewis, D. A. & Levitt, P. *Mol. Psychiatry* **6**, 293–301 (2001).
11. Oliveira-Dos-Santos, A. J. *et al.* *Proc. Natl Acad. Sci. USA* **97**, 12272–12277 (2000).
12. Heximer, S. P. *et al.* *J. Clin. Invest.* **111**, 445–452 (2003).

Semiconductor physics

Relativity on a chip

Michael E. Flatté

The manipulation of electronic spins — 'spintronics' — might be the basis of future device technology. A subtle relativistic effect offers a way to flip spins inside a semiconducting material.

Two great revolutions in physics in the early twentieth century — quantum mechanics and Einstein's relativity — altered how we understand electromagnetism and motion. In the semiconductor-based technologies of modern-day electronics, the motion of electrons and their interactions are governed by quantum mechanics; the rules of relativity are mostly peripheral. But relativistic phenomena are not outside the realm of common experience. For instance, the magnetic force between two wires carrying moving charges (electric current) arises because moving charges in one wire 'perceive' a relativistic length contraction between moving charges in the other¹. However, semiconductor devices that rely on fundamentally relativistic principles are rare.

Kato *et al.*² (as they report on page 50 of this issue) have exploited relativity to manipulate the spin of electrons in a simple thin slab of semiconductor crystal. Spin is a

property of every electron, causing it to behave like a small bar magnet; its 'north' and 'south' can be reoriented by magnetic fields, but are unaffected by electric fields. Kato *et al.*, however, have succeeded in reorienting electron spins without applying a magnetic field to their crystal. The trick is to use intrinsic magnetic fields that arise through the effects of relativity.

According to the theory of relativity, the energies of electrons moving in a solid change slightly, through the so-called spin-orbit interaction³. In essence, the purely electric fields around the static atoms of the solid crystal are 'seen' by moving electrons as magnetic fields⁴. So, although the spin of a motionless electron cannot be affected by the atomic electric fields, the spin of a moving electron can. The perceived magnetic field causes the electron spin to precess, like a spinning-top, keeping its direction at the same angle to the magnetic field. If the field is

strong enough and oriented perpendicular to the spin, the spin can be induced to turn right around and point in the opposite direction.

The electrons in a semiconductor crystal are moving in all directions at once, even though their spins may be aligned in a single direction (the population is 'spin-polarized'). So the average effect of those relativistic magnetic fields does not produce a coherent reorientation of the spin of the electron population. The fields do, however, dominate the incoherent relaxation of the spin polarization^{5,6}. To obtain a preferred rotation axis there must be more electrons moving in one direction than in the other⁷ — in other words, a flow of electrical current.

Kato *et al.*² applied a voltage to a thin slab of gallium arsenide (GaAs) crystal through small contacts placed 300 μm apart, forcing the spin-polarized electrons to move with a constant average velocity. Even with this current flow, the size of the spin precession in crystals such as GaAs is usually extremely small: a coincidental mathematical cancellation means that the magnitude of the effective magnetic field is proportional not simply to the electron speed (typically around one kilometre per second), but to the cube of the electron speed⁸. But Kato *et al.* overcame the coincidental cancellation by introducing a 'stressor' layer of aluminium gallium arsenide underneath the 2- μm -thick GaAs film. The atoms in the stressor layer are slightly more widely spaced than those in GaAs and force them farther apart, straining the GaAs film. As a result, the effective magnetic fields are orders of magnitude larger.

For modest electric-field values of 100 V cm^{-1} , Kato *et al.* saw the electron spins rotate through a full turn and a half (an angle of 3π) as they travelled 60 μm through the GaAs layer. Such flipping of the electron spin direction suggests that this will be a useful approach in semiconductor technology, to control and manipulate the information contained in the average spin polarization of the electrons in the material.

In fact, Kato *et al.*² have removed several roadblocks in the path of semiconductor 'spintronic' commercialization at a stroke — such as the need for an applied magnetic field or for the use of exotic materials. Schemes for small-scale engineering of strained structures similar to that used by the authors are well established. So there is great promise for the application of relativistic magnetic fields in electronic devices.

There is perhaps one drawback — this experiment was performed at low temperature (5 K). But there is no fundamental reason why it should not work at room temperature in other materials. There are already candidates, such as gallium antimonide and indium antimonide, that are known to exhibit much larger relativistic magnetic fields for moving electrons. Perhaps relativity is at last set to become

a key component in the semiconductor-device toolbox.

Michael E. Flatté is in the Optical Science and Technology Center and the Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52242, USA.

e-mail: michael_flatte@mailaps.org

1. Purcell, E. M. *Electricity and Magnetism* (McGraw-Hill, New York, 1984).

2. Kato, Y., Myers, R. C., Gossard, A. C. & Awschalom, D. D. *Nature* **427**, 50–53 (2004).
3. Merzbacher, E. *Quantum Mechanics* (Wiley, New York, 1970).
4. Landau, L. D. & Lifshitz, E. M. *The Classical Theory of Fields* (Pergamon, New York, 1975).
5. Yafet, Y. *Solid State Phys.* **14**, 1–98 (1963).
6. D'yakonov, M. I. & Perel', V. I. *Sov. Phys. JETP* **33**, 1053–1059 (1971).
7. Kalevich, V. K. & Korenov, V. L. *JETP Lett.* **52**, 230–235 (1990).
8. Dresselhaus, G. *Phys. Rev.* **100**, 580–586 (1955).

Palaeontology

Chinese lantern for early primates

Robert D. Martin

A fossil skull from China, dating to 55 million years ago, provides much-needed substantial evidence of early primates in Asia. Interpretation of the creature's eye size and activity pattern will spark debate.

The first fossil evidence for 'primates of modern aspect' — the euprimates — appears abruptly in the northern continents at the beginning of the Eocene, about 55 million years ago. Eocene primates are well documented from sites in North America and Europe, but specimens from Asia have been scarce and fragmentary. That situation changes with the report by Ni *et al.* of a well-preserved partial primate skull, with lower jaws, from China (page 65

of this issue¹). The authors allocate the fossil to a new species of the genus *Teilhardina*, which was originally established for fragmentary material from Belgium².

The new skull has several euprimate characteristics, notably a bony ring around each eye-socket (orbit), forward rotation of the orbits and a relatively large braincase. But the skull is very small, with an overall length of about 25 mm, and the animal's body mass is estimated as 28 g — less than in any modern

primate species. As expected for one of the earliest known euprimates, the dentition is relatively primitive. The upper and lower jaws both contain four premolars on each side (the maximum known for any primate) and the front premolars are not as reduced as in other *Teilhardina* species. Certain dental features, and its small body size, suggest that the animal was an insectivore.

The bigger picture is outlined in Fig. 1, which reflects the widely accepted view that there is a basic dichotomy in the primate evolutionary tree — one lineage leading to modern lemurs and lorises (strepsirrhines), the other to tarsiers, monkeys, apes and humans (haplorhines)^{3–5}. All or most Eocene primates are allocated to two major groups, lemuroids and tarsoids. These are commonly linked to modern strepsirrhines and haplorhines, respectively (although both could possibly have arisen from an independent northern radiation of primates^{6,7}). *Teilhardina* is included among tarsoids as a potential relative of haplorhines.

As well as the new species (*T. asiatica*), and the originally described *T. belgica*², the genus *Teilhardina* currently includes five other species from North America. On this basis, then, *Teilhardina* seemingly had a remarkably wide distribution. Using characters derived from the new skull, Ni *et al.*¹ repeated an earlier analysis of primate relationships⁵. The results confirm the basic primate dichotomy, and are striking for *Teilhardina* itself. *Teilhardina asiatica* and *T. belgica* together constitute the earliest offshoot from the haplorhine side of the tree, branching off before the split between other tarsoids and haplorhines (Fig. 1). By contrast, *T. americana* (the only American species considered) lies within the cluster containing all other tarsoids, reflecting a pronounced separation from *T. asiatica* and *T. belgica*. Inclusion of *T. americana* and other closely related North American species in the genus *Teilhardina* is therefore questionable, as is the apparent extensive distribution of the genus.

Another unusual feature of *T. asiatica* is the small size of the orbits relative to skull length. Analysis of the scaling of orbit size using a published data set⁸ indicates that this species falls on the best-fit line for modern diurnal primates, which generally have relatively smaller orbits than nocturnal species. From this, Ni *et al.*¹ infer that *T. asiatica* was diurnal. They go on to interpret the evolution of nocturnal versus diurnal habits among primates, and conclude that the last common ancestor of euprimates was a diurnal, visually oriented predator.

This is an important issue, one on which I have to disagree with the authors on both statistical and biological grounds. Prediction from regression lines beyond the range of the original data is suspect, and *T. asiatica* is smaller than any modern primate. In fact, Ni and colleagues excluded all large-bodied

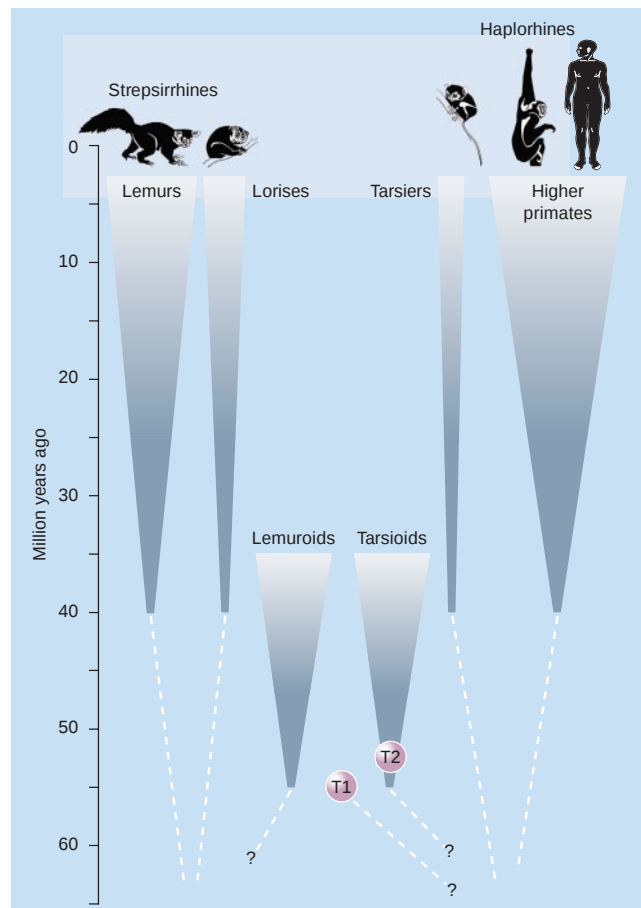


Figure 1 **Primate evolution in outline.** This tree incorporates the results of the analysis by Ni *et al.*¹ (see Fig. 3 on page 67; lemuroids are formally known as Adapiformes, and tarsoids as Omomyiformes). Along with *Teilhardina belgica*, the new species *T. asiatica* branches away first on the haplorhine side of the tree (T1). By contrast, *T. americana* (T2) is nested within the other Eocene tarsoids, calling into doubt its place in the genus *Teilhardina*. (Primate icons drawn by Lucrezia Beerli-Bieler.)

primates from the original data set, so altering the slope of the best-fit line for diurnal primates and strengthening the inference that *T. asiatica* was diurnal. Furthermore, no account was taken of the potential problem of 'phylogenetic inertia'⁹: closely related species may not be statistically independent, so estimates of probability may be open to doubt.

Biologically, one cannot assume that early primates (particularly if unusually small in size) showed the same functional patterns as modern primates — which themselves are very variable. Given that ancestral primates descended from ancestral mammals with smaller eyes, early primates presumably showed only moderate enlargement of their eyes, regardless of their habits. In early primates, the brain was less than half the size of the brains of their modern relatives³. So processing of visual inputs was probably more rudimentary, perhaps explaining why some Eocene lemuroids had far smaller orbits than modern diurnal lemurs.

Another source of information comes from a study by Kay and Kirk⁶. To increase visual sensitivity in dim light, nocturnal species show marked retinal summation of inputs from the photoreceptors, resulting in a narrower optic nerve. This is reflected by a narrower opening (foramen) for the optic nerve in the back of the orbit. For living primates, Kay and Kirk found a good match between activity pattern and relative size of the foramen. However, all Eocene primates have a small foramen, regardless of whether they have relatively small orbits (suggesting diurnal vision) or large orbits (indicating nocturnal habits). Thus, visual adaptations in early primates were clearly different from those in modern primates. Finally, a feature on the snout of *T. asiatica*, the infraorbital foramen, is markedly larger than in modern primates. The size of this foramen — large in primitive nocturnal mammals, typically

reduced in diurnal mammals — indicates the degree of development of tactile whiskers for non-visual orientation.

I believe, then, that we remain in the dark with respect to the activity pattern of *T. asiatica*. Regardless of that, however, Ni and colleagues' discovery is notable not only for its implications for primate systematics but also from a biogeographical perspective. Until recently, it was widely held that direct migration of mammals between Asia and Europe around 55 million years ago was ruled out by a transcontinental marine barrier. The landmass of Eurasia was largely or completely split down the middle by a combination of the Western Siberian Obik Sea to the north and the Turgai Straits to the south. So it was suggested that the only possibility for interchange of mammals between Asia and Europe was indirect migration across the Bering Straits, through North America and across the Greenland landbridge, or vice versa. However, the presence of closely related *Teilhardina* species in China and Belgium (but not in America) in the early Eocene adds to evidence that migration between Asia and Europe did not necessarily involve such a roundabout route¹⁰.

Robert D. Martin is in Academic Affairs, The Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605-2496, USA.

e-mail: rdmartin@fieldmuseum.org

1. Ni, X., Wang, Y., Hu, Y. & Li, C. *Nature* **427**, 65–68 (2004).
2. Simpson, G. G. *Bull. Am. Mus. Nat. Hist.* **77**, 185–212 (1940).
3. Martin, R. D. *Primate Origins and Evolution: A Phylogenetic Reconstruction* (Princeton Univ. Press, 1990).
4. Fleagle, J. G. *Primate Adaptation and Evolution*, 2nd edn (Academic, New York, 1999).
5. Ross, C. F., Williams, B. & Kay, R. F. *J. Hum. Evol.* **35**, 221–306 (1998).
6. Martin, R. D. *Nature* **363**, 223–234 (1993).
7. Ross, C. F. *J. Hum. Evol.* **45**, 195–201 (2003).
8. Kay, R. F. & Kirk, E. C. *Am. J. Phys. Anthropol.* **113**, 235–262 (2000).
9. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, 1991).
10. Hooker, J. J. & Dashzeweg, D. *Geol. Soc. Am. Spec. Pap.* **369**, 479–500 (2003).



100 YEARS AGO

Under the title of "The Case for Vaccination," Mr. C. E. A. Winslow gives an admirable survey of the statistical data in favour of the efficacy of vaccination (*Science*, July 24, p. 101). It points out that a single vaccination greatly reduces the probability of an attack of small-pox, postpones it to a later period of life, and renders it less dangerous if it does ensue. To ensure absolute protection revaccination is required. During the small-pox epidemic of 1871, of 734 nurses and attendants in the Metropolitan Asylums Board Hospitals 79 were survivors from small-pox attack, and escaped infection; 645 were revaccinated on entrance, and all escaped; 10 were not revaccinated, and all took small-pox. Mr. Winslow concludes, "if statistics ever proved anything, those quoted prove the protective influence of vaccination." From *Nature* 31 December 1903.

50 YEARS AGO

Scientific centenaries in 1954. The modern use of a surname is so well established that it is worth recalling that, as late as the fifteenth century, hereditary surnames were by no means general, it being the custom to add to the baptismal name the place of the owner's birth or residence, or his occupation or some peculiarity that would identify him. Thus we find that Hermann, an eleventh-century monk, was known as Hermann of Reichenau, after the abbey in Switzerland where he spent most of his life, or Hermann the Cripple, on account of the paralysis which had afflicted him from youth. Hermann, born in 1013, has been described as one of the most learned men of his age, and interests us particularly because he was devoted to the study of mathematics and astronomy... He died in 1054, at the castle of his father, Count Wolferad of Alshausen... The year 1954 marks the five hundredth anniversary of the birth of the Italian astronomer Domenico Maria di Novara... Passing from Italy to Germany, we note the death in 1554 of Jerome Bock, the herbalist... His great herbal, "New Kreütter Büch", which appeared first in 1539, is remarkable because Bock was the first botanist to attempt descriptions, from direct observation, which would render illustrations unnecessary. To obtain his material he made long journeys on foot, dressed as a peasant, to avoid undue attention. Illness and misfortune clouded his later life, and he died when only fifty-six, predeceased by eight of his ten children. From *Nature* 2 January 1954.

Oceanography

The southern supplier

Joachim Ribbe

Physical processes in the Southern Ocean largely control nutrient distribution in the global marine environment, a finding that further highlights the influence of this oceanic region on Earth's climate.

To understand how climate change comes about, and what the future may hold, we need to untangle the linkages between ocean circulation and the productivity of phytoplankton. Productivity depends on nutrient availability in the ocean and, as phytoplankton are leading players in the global carbon cycle, they partly determine levels of the greenhouse gas carbon dioxide in the ocean and atmosphere.

On page 56 of this issue, Sarmiento *et al.*¹

argue that the Southern Ocean controls the distribution of nutrients in most of the upper ocean throughout the world. It does so through the formation of nutrient-rich water masses, which spread throughout the Southern Hemisphere and into the North Atlantic. A similar process occurs in the North Pacific, but makes a smaller contribution to nutrient availability.

The formation of water masses within defined geographical regions links the global

ocean and the atmosphere, and is a central mechanism of oceanic control of climate. They constitute a repository for heat, fresh water and gases such as carbon dioxide, all of which are exchanged at the ocean–atmosphere interface. On the largest scale, the water masses spread and fill the ocean basins. These huge bodies of water are also the engines of large-scale ocean circulation, one that is primarily driven by so-called North Atlantic Deep Water (NADW). As its name suggests, this is water that forms at high latitudes in the North Atlantic, sinks to depth and flows southwards.

Since the initial recognition of this circulation², oceanographers have busied themselves with finding out how NADW is resupplied to the Northern Hemisphere. A complex pattern of circulation pathways has become evident, with the Southern Ocean (Fig. 1) apparently having an important role in the production and interchange of water masses^{3,4}. Subantarctic Mode Water (SAMW) is a large water mass created by exchange of heat and fresh water with the atmosphere over much of the Southern Ocean. It sinks below the ocean surface^{5,6} and moves northwards at depths of about 200–600 m. These regions of the Southern Ocean can be thought of as windows to the deep ocean that allow water of particular heat, freshwater and nutrient content to enter the subsurface ocean.

The surface regions where SAMW is formed are characterized by low levels of silicic acid and high concentrations of nitrate. Sarmiento *et al.*¹ apply a newly designed 'conservative' tracer that captures this nutrient signature as a characteristic of SAMW. Their analysis of its distribution shows that SAMW reaches most of the world's upper ocean — that is, much of the global marine environment receives nutrients from the surface of the Southern Ocean. Another water mass, known as North Pacific Intermediate Water, has a subordinate role only: its nutrient delivery is limited to the upper ocean of the North Pacific.

Sarmiento *et al.*¹ also performed several computational experiments. They used a model of the global ocean that couples physical and biological processes, and came up with a reason why diatoms, a major component of the phytoplankton, do not reach their full productive potential in much of the global ocean. By varying the strength of the nutrient source in their models, Sarmiento *et al.* found that the ratio of silicic acid to nitrate in SAMW is less than ideal for diatom growth, and is therefore a primary cause of diatoms' widespread low productivity.

The conclusion that a physical process, and one operating in only a tiny part of the world's oceans, has such a huge influence on productivity is startling in itself. But there are more lessons to be learned from the new work¹.

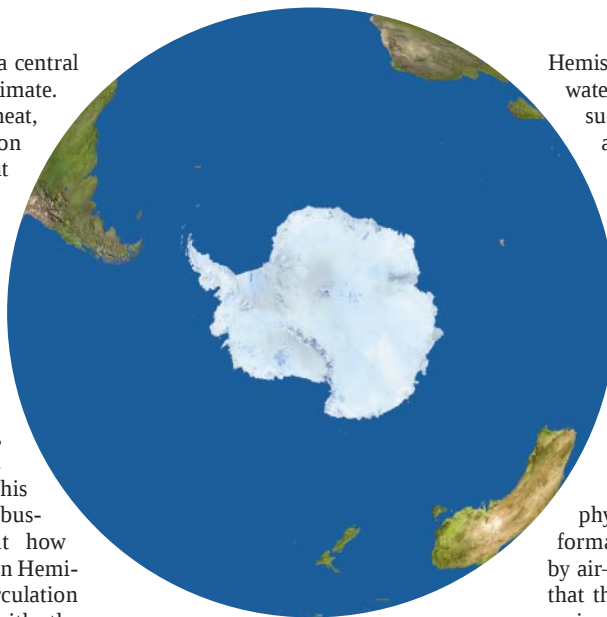


Figure 1 Well connected. The Southern Ocean surrounds Antarctica but has a global influence on marine productivity.

First, use of the new tracer as a diagnostic tool allows further insight into the dynamics of present-day ocean circulation. In particular, it has revealed that Antarctic surface water, driven northwards by the winds, contributes to the characteristics of SAMW. This finding supports earlier studies of this water mass^{6,7} and highlights the more general point that models of the climate system have to capture accurately the physical processes through which water masses form.

Second, the simulations identify a possible deficiency in representing mixing processes in the North Pacific. Models of the climate system may have to include the possible effect of tides on water-mass formation globally, which in turn has implications for the global influence exercised by the outflow characteristics of NADW.

Third, the SAMW pathway does more than provide nutrients for the upper branch of the ocean circulation. In the Southern

Hemisphere it also supplies heat and fresh water to the thermocline — the layer of sudden temperature change that occurs at varying depth below the surface and effectively separates the upper, productive part of the ocean from the deeper waters. The authors do not stress this point, but it is highly topical, given its bearing on our understanding of climate variability in the tropics. At the moment there is vigorous debate over the question of oceanic 'teleconnection' of high and mid-latitudes with low latitudes, which may drive year-to-year climate variability in the tropics⁸.

As to the climate connection, the physical processes that lead to water-mass formation in the Southern Ocean are driven by air–sea interactions. The indications are that the Southern Ocean is one of the few regions of the global ocean where the atmospheric consequences of climate change enter the ocean⁹. Any perturbation of the formation of SAMW, and subsequently of ocean circulation, is likely to have a dramatic impact on global marine productivity. That will affect the levels of carbon dioxide in the atmosphere, which in turn are linked to global temperature changes¹⁰. The overall conclusion is that the Southern Ocean has as powerful an influence on climate change as the North Atlantic. Difficult though it may be, it would pay to monitor conditions there. ■

Joachim Ribbe is in the Department of Biological and Physical Sciences, University of Southern Queensland, Toowoomba, Queensland 4350, Australia.

e-mail: Joachim.Ribbe@usq.edu.au

1. Sarmiento, J. L., Gruber, N., Brzezinski, M. A. & Dunne, J. P. *Nature* **427**, 56–60 (2004).
2. Broecker, W. S. *Oceanography* **4**, 79–89 (1991).
3. MacDonald, A. M. & Wunsch, C. *Nature* **382**, 436–439 (1996).
4. Rahmstorf, S. *Nature* **419**, 207–214 (2002).
5. Ribbe, J. *Tellus Ser. A* **51**, 517–525 (1999).
6. Sørensen, J. *et al.* *J. Phys. Oceanogr.* **31**, 3295–3311 (2001).
7. Rintoul, S. R. & England, M. H. *J. Phys. Oceanogr.* **32**, 1308–1321 (2002).
8. Gu, D. & Philander, S. G. H. *Science* **275**, 805–807 (1997).
9. Banks, H. T. *et al.* *Geophys. Res. Lett.* **27**, 2961–2964 (2000).
10. Petit, J. R. *et al.* *Nature* **399**, 429–436 (1999).

Cell biology

Shape-shifting protein channel

Jordi Benach and John F. Hunt

Newly made proteins are moved across cellular membranes through a protein channel. The crystal structure of this channel is now revealed and confirms expectations that it must change shape to allow proteins to pass.

All cellular proteins are synthesized in the body of the cell, the cytosol. But many of them must then be transported through phospholipid membranes to reach their final destinations, which might be intracellular compartments or even, following secretion, outside the cell^{1–5}. The

molecular mechanism of this 'translocation' process has been the subject of elegant biochemical^{1,4–9}, genetic³ and biophysical^{10–12} studies. This body of work has shown that the main secretion pathway in all kingdoms of life involves a heterotrimeric protein complex, or translocase^{4,5}, which forms a

Atmospheric pollution

The veil of two cities

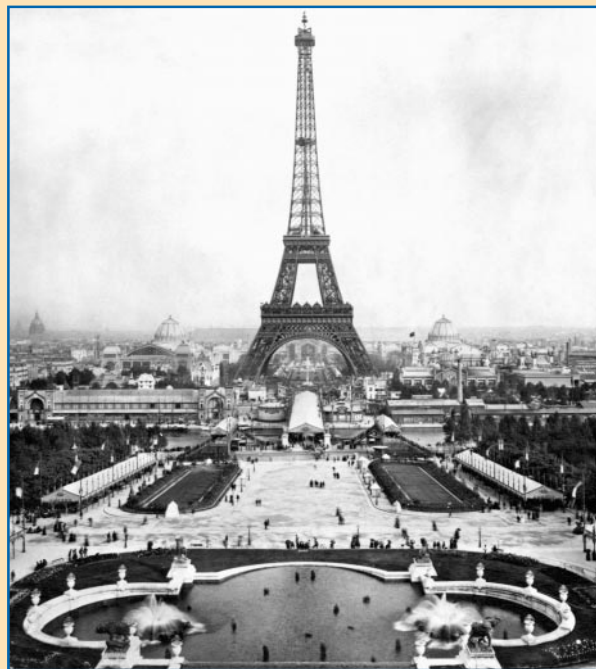
In 1752, Benjamin Franklin sent a metal key on its famous kite flight to demonstrate the electrical nature of lightning. The study of atmospheric electricity subsequently burgeoned, and writing in *Atmospheric Environment* (37, 5319–5324; 2003) R. G. Harrison and K. L. Aplin describe how they have mined records of one aspect of later historical research. They show that electrical measurements made at the Eiffel Tower in the 1890s can be used to estimate the levels of Parisian smoke pollution at that time.

The atmosphere's electrical system results from a combination of thunderstorm activity and the slight ionization of air. The upshot is an electrical potential difference between the ionosphere (at an altitude of some 60 km and more) and the Earth's surface, which causes a small current to flow. The potential gradient close to the ground is an indicator of the electrical state of the atmosphere, and in the nineteenth century it was commonly measured in several European cities.

In clean air, the potential gradient has a distinctive diurnal cycle known as the Carnegie curve. Local aerosol pollution alters this

gradient. So if that effect can be separated from the influence of the global electrical circuit at atmospherically clean sites, it becomes possible to infer pollution levels from potential-gradient measurements. But that is easier said than done.

To analyse electrical measurements made at the top of the Eiffel Tower, Harrison and Aplin applied a relationship between smoke pollution and potential gradient that had emerged from their earlier historical study for Kew, near London, using data from 1863. But they also needed an absolute calibration that was specific for Paris. For this, the authors conducted a modelling study that enabled them to identify the time of day when the top of the Eiffel Tower was in clean air and when it was in the urban boundary layer — the atmospheric layer affected by urban pollution. At times when the tower was in clean air, the observed electrical variations can be attributed to the Carnegie curve alone, allowing an absolute calibration of the electrical data based on the tabulated Carnegie values. At times when it was in the urban boundary



layer, the higher potential gradient signals the presence of smoke pollution.

The approach allowed Harrison and Aplin to estimate smoke pollution at both the top and the bottom of the Eiffel Tower in the 1890s. They calculate that pollution at ground level had an autumnal daily peak of $60 \pm 30 \mu\text{g m}^{-3}$. Their

earlier study gives a value of $170 \pm 50 \mu\text{g m}^{-3}$ for London. In the 1860s the UK Clean Air Act lay almost 100 years in the future. So in the late nineteenth century it is likely that, as a place to live, Paris had the edge over its English counterpart in more than just cancan and cabaret.

Juliane Mössinger

tightly controlled conduit that allows newly made proteins to pass through membranes before the proteins fold into their functional shape.

On page 36 of this issue, van den Berg and colleagues¹³ present the X-ray crystal structure of the translocase (the SecY $\text{E}\beta$ complex) from the single-celled archaeon *Methanococcus jannaschii*, providing an initial view of the atomic architecture of this universally conserved channel. This structure could be considered the latest triumph of genomics, because the choice of the best translocase for structure determination was made on the basis of empirical examination of the expression, purification and crystallization properties of a range of translocases from organisms with fully sequenced genomes.

Almost 30 years ago, as a corollary to his 'signal hypothesis', Günter Blobel^{1,2} proposed that the translocation of proteins across membranes would occur through a proteinaceous channel of the kind now crystallized. Blobel's research at that time had shown that newly made proteins ('preproteins') that are

targeted for export from the cytosol have an extension at one end, called a signal peptide, that is removed during passage through the membrane¹. The hypothesis proposed that different types of extension would function as signals, directing newly made proteins to different membrane-bounded compartments^{1,2}, and the importance of this insight was rapidly accepted. The proposal that protein translocation occurs through a protein channel remained controversial for some time, until later experiments^{3–7,10} confirmed it.

Ion channels have received considerable attention because they show remarkable specificity in allowing one kind of ion through while preventing the passage of very similar molecular species. The protein-translocation channel faces what could be considered a more daunting task, in that it must allow the passage of chemically and sterically varied substrates — representing any segment from a translocating protein — without compromising the permeability barrier of the membrane^{2,4–6,11}. Blobel's original answer to this conundrum was that

translocation would occur at the same time that the protein was being made¹ (co-translationally), with a tight seal between the protein-synthesis machinery (ribosomes) and the pore of the translocation channel maintaining osmotic integrity. But biochemical studies have shown that translocation occurs at least partially post-translationally^{4–7}, and cryo-electron-microscopic studies show a gap of some 15 Å separating the ribosomal exit site from the translocase in detergent-solubilized samples performing co-translational translocation¹⁰ (as in Fig. 1, overleaf). So the pore of the channel seems likely to have variable but controlled conformational properties, expanding just enough to accommodate larger protein segments without compromising the integrity of the osmotic seal.

The crystal structure of the SecY $\text{E}\beta$ channel presented by van den Berg *et al.*¹³ provides an initial atomic-level glimpse of this shape-shifting channel. The authors advance structural arguments to support the hypothesis that the channel's pore is located at the centre of a single SecY $\text{E}\beta$ heterotrimer.

Previous work had found that biochemically well-characterized translocases, such as that from *Escherichia coli*, purify as higher-order multimers^{4,5,10}, which led to the suggestion that the pore would be located at the interface between several heterotrimers. This possibility was supported by biophysical studies suggesting that the pore is too large to be accommodated in a single heterotrimer¹¹. Although these observations suggest that several heterotrimers might fuse to form a larger channel under certain circumstances, the hypothesis that individual heterotrimers are active in protein translocation is supported by biochemical studies of the *E. coli* complex^{7,8}.

The SecYE β structure¹³ seems to be in a closed state — not surprisingly, given that the complex was crystallized in the absence of a preprotein substrate. With support from biochemical and genetic results, van den Berg *et al.* propose that a short α -helix (a canonical structural feature in proteins) that plugs the exit from the channel will move out of the way during protein translocation. But removal of this plug would expose a pore whose molecular surface, at its narrowest point, has a diameter of only 3 Å (Fig. 1). The minimal requirement for channel function would be the ability to pass an arbitrary preprotein segment in an extended conformation, which would have a maximal width of around 12 Å. Moreover, several observations suggest that the channel can accommodate larger transport substrates, including α -helices destined to adopt a transmembrane conformation^{2,5,9}, which have a diameter of some 14 Å. So the channel must dynamically expand and contract in a way that is coordinated with the passage of preprotein segments.

Surprisingly, van den Berg *et al.* find that the limiting constriction — the 3-Å pore in the channel — is lined by a ring of inflexible isoleucine amino acids. This suggests that the inferred dynamic resizing of the channel must be mediated by diaphragm-like movements of the constituent transmembrane α -helices of the SecYE β heterotrimer, rather than by changes in side-chain conformation. Movements of the α -helices in SecYE β that are considerably larger than those that allow preprotein passage are also inferred from evidence that hydrophobic transmembrane α -helices in translocating preproteins can be released directly into the membrane bilayer when their presence is sensed in the translocation channel⁹. Van den Berg *et al.* propose a pathway for this sideways release, involving disruption of the interface between two pseudosymmetric halves of the SecY subunit of SecYE β , leading to opening of the central channel to the membrane.

Future experiments will need to explore how these inferred conformational changes are controlled by, and coupled to, preprotein translocation. In addition to regulation by

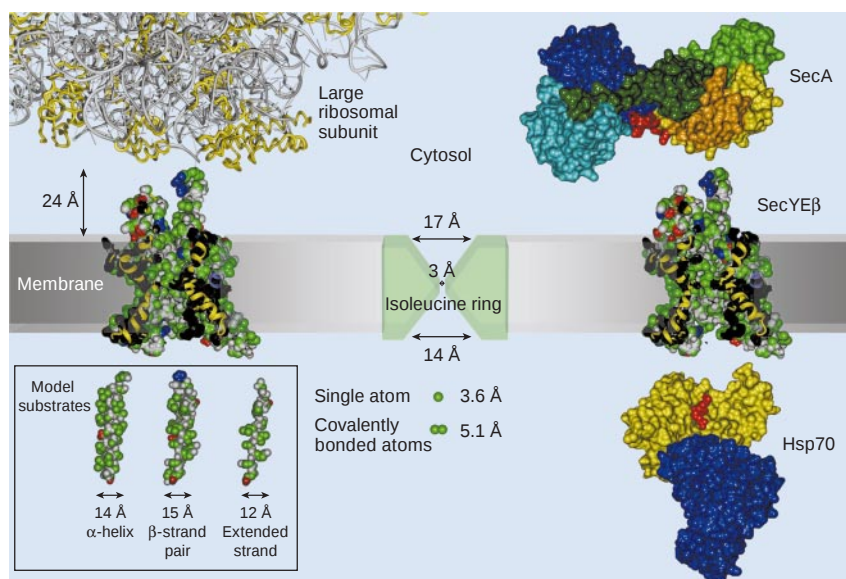


Figure 1 The SecYE β channel¹³, with some of its binding partners and model substrates shown at the same scale. The channel transports newly made proteins across cell membranes. Two cut-away views through its centre are shown at left and right, with the α -helix that blocks the exit channel moved into the proposed 'open' position¹³. Hydrophobic amino-acid side chains are in green, acidic groups in red, basic groups in blue, and other atoms in grey; yellow ribbons denote α -helices of SecY inside the surface. Space-filling representations of model substrates use the same colours. The channel's molecular dimensions are shown in the middle. The large ribosomal subunit synthesizes proteins; RNA is in grey and the subunit's protein backbone in yellow, with its exit channel positioned roughly as seen in ref. 10. A surface representation of an Hsp70 protein related to BIP shows the peptide-binding domain in yellow, bound peptide in red and ATP-hydrolysing domain in blue (with relative orientation roughly as in ref. 14); BIP prevents preproteins from slipping back into the channel. A surface representation of SecA from *Bacillus subtilis* shows the first (blue) and second (cyan) ATP/ADP-binding folds, amino-terminal (yellow) and carboxy-terminal (orange) preprotein-crosslinking domains, scaffold domain (dark green), wing domain (light green), and carboxy-terminal linker (red)¹⁵. This protein helps to push preproteins into the channel. The figure was created with DINO¹⁶. Protein Data Bank accession numbers are available from us on request.

translocating transmembrane α -helices, van den Berg *et al.* propose a model in which the channel's dynamics are modulated directly by the signal peptide of the preprotein being transported. Some of the required conformational changes could also be coupled to the protonmotive force (a transmembrane gradient in hydrogen-ion concentration and/or electrical potential), which enhances the efficiency of protein translocation in bacteria⁶.

Both these possibilities need to be addressed, as does the manner in which the channel is regulated by its known protein partners (Fig. 1). In bacteria, for instance, the SecA protein uses energy from ATP hydrolysis to push preprotein segments through the channel^{4,6,7}. How does the binding of SecA to the channel's cytosolic surface control the expansion and contraction that allows preproteins to pass while maintaining osmotic integrity? Finally, in eukaryotic (nucleated) cells, the Hsp70-family chaperone BIP is believed to act like a ratchet that prevents translocated preprotein segments from slipping back into the channel. Given that BIP and the channel interact¹², BIP might also regulate channel dynamics. The crystal

structure of the SecYE β channel¹³ will allow mechanistic issues such as these to be addressed with greater structural specificity and sophistication.

Jordi Benach and John F. Hunt are in the Department of Biological Sciences, 702A Fairchild Center, MC2434, Columbia University, New York, New York 10027, USA.

e-mail: hunt@sid.bio.columbia.edu

1. Blobel, G. & Dobberstein, B. *J. Cell Biol.* **67**, 835–851 (1975).
2. Blobel, G. *Proc. Natl. Acad. Sci. USA* **77**, 1496–1500 (1980).
3. Schatz, P. J. & Beckwith, J. *Annu. Rev. Genet.* **24**, 215–248 (1990).
4. de Keyser, J., van der Does, C. & Driessen, A. J. *Cell. Mol. Life Sci.* **60**, 2034–2052 (2003).
5. Rapoport, T. A., Jungnickel, B. & Kutay, U. *Annu. Rev. Biochem.* **65**, 271–303 (1996).
6. Schiebel, E., Driessen, A. J., Hartl, F. U. & Wickner, W. *Cell* **64**, 927–939 (1991).
7. Yahr, T. L. & Wickner, W. T. *EMBO J.* **19**, 4393–4401 (2000).
8. Duong, F. *EMBO J.* **22**, 4375–4384 (2003).
9. Do, H., Falcone, D., Lin, J., Andrews, D. W. & Johnson, A. E. *Cell* **85**, 369–378 (1996).
10. Beckmann, R. *et al. Cell* **107**, 361–372 (2001).
11. Hamman, B. D., Chen, J. C., Johnson, E. E. & Johnson, A. E. *Cell* **89**, 535–544 (1997).
12. Hamman, B. D., Hendershot, L. M. & Johnson, A. E. *Cell* **92**, 747–758 (1998).
13. van den Berg, B. *et al. Nature* **427**, 36–44 (2004).
14. Zhu, X. *et al. Science* **272**, 1606–1614 (1996).
15. Hunt, J. F. *et al. Science* **297**, 2018–2026 (2002).
16. www.dino3d.org

Immunology

Bowel disease and bone loss

Immunity **19**, 849–861 (2003)

Certain types of bone disorder and inflammatory bowel disease probably occur when the body mounts an immune response against itself. A subset of immune cells, known as CD4-expressing T cells, is overactive in these circumstances — but their exact role in disease is unclear.

When mice are engineered to lack a key regulator of these T cells, the cells become hyperactive and the mice develop colon inflammation and osteoporosis-like bone wasting. A. J. Ashcroft *et al.* now report that the animals' condition is caused by a protein called RANKL. They find that the hyperactive T cells produce too much of the protein, which then contributes to bone breakdown and bowel inflammation. When the mice are treated with osteoprotegerin — a protein that interferes with RANKL's binding to its receptor — bone loss is reversed and colon inflammation lessens. The study suggests that some bone diseases and intestinal problems may have a common cause.

Helen R. Pilcher

Networks

Special friends

Phys. Rev. Lett. **91**, 247901 (2003)

In epidemiology, topology matters: the structure of a network in which an infectious disease spreads can have a big influence on how easy it is to arrest the disease. Some social networks seem to be 'scale-free' — bound together by just a few very highly connected individuals (hubs). An ideal immunization strategy for, say, preventing epidemics of sexually transmitted disease in such a network would be to find and immunize these hubs. But that can be extremely hard to do. Random immunization, on the other hand, is relatively ineffective, for there is only a low chance of eliminating hubs short of immunizing almost everyone, which is costly. The same principles hold true for containing computer viruses, as electronic networks seem also to be scale-free.

Reuven Cohen and colleagues now propose an alternative strategy. On the basis of mathematical modelling, they suggest that immunizing just a small fraction of the acquaintances, chosen at random, of randomly selected individuals will dramatically reduce the epidemic threshold — the fraction of all individuals in the network that must be immunized to

prevent an epidemic. Although this strategy might sound little different from random immunization, it exploits the fact that acquaintances tend on average to have more connections than individuals selected at random: the mere fact of being a 'friend' makes them special.

Philip Ball

Evolutionary biology

Sexual conflict costs

Proc. R. Soc. Lond. B doi:10.1098/rspb.2003.2588

For a female dung fly, having a choice of mates comes at a price. Females selected for the rough-and-tumble of multiple mating do worse than monogamous strains when they then mate only once, say Oliver Y. Martin and his colleagues.

The researchers subjected dung flies, *Scathophaga stercoraria*, to ten generations of artificial selection. One set of females was allowed multiple matings (polyandry), while monogamy was forced on another set. Martin *et al.* then looked at the fitness of each line when mated just once: they found that polyandrous females had shorter lives, and fewer eggs and offspring.

When females mate with many males, as happens in most insects, the conflict between the sexes increases. Females wish to control whose sperm they use to fertilize their eggs, but evolution will drive males to try and manipulate females into using their sperm. Removing this competition reveals the costs that adapting to fight the battle imposes on females. It is not known what reduces the fitness of polyandrous females, although it has shown that their immune systems are weaker than those of their monogamous counterparts.

John Whitfield

Chemical detection

Odour eater

Chem. Commun. 2970–2971 (2003)

As Agatha Christie well knew, the smell of bitter almonds can indicate cyanide in the air and death not far away. In reality, industrial workers who might be exposed to

this toxic gas need appropriate respiratory protection. Michael J. Hudson *et al.* describe a new material that could be used in such apparatus as a filter for hydrogen cyanide gas, and which has many practical benefits compared with conventional activated-carbon systems.

The material is an ordered mesoporous silica that has been doped with sodium peroxydisulfate. It is simple to synthesize and can filter out hydrogen cyanide for longer than can a range of other porous materials. The doped silica also functions in an aqueous environment, whereas existing filter materials tend to fail at high humidity. In addition, it does not contain the carcinogenic chromate ions that are present in some conventional materials; nor does it produce cyanogen, another poisonous molecule.

Hudson *et al.* admit that further development is needed, in particular to improve the stability of the material. But its efficacy as a cyanide filter could well have fooled Miss Marple.

Rosamund Daw

Cancer

Creating instability

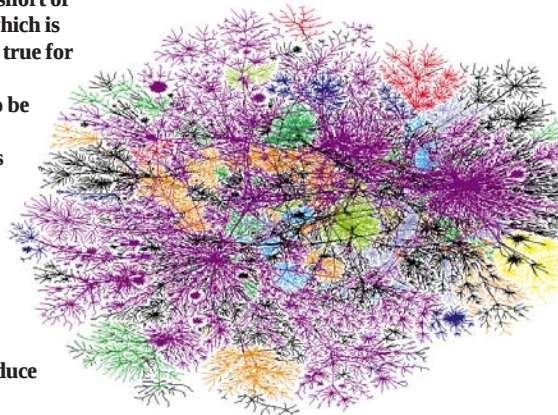
J. Cell Biol. **163**, 949–961 (2003)

Colorectal cancer is one of the most common cancers worldwide. Although the cause remains unclear, tumour progression may result from the abnormal gain or loss of chromosomes during cell division — a process called chromosome instability. Rebecca A. Green and Kenneth B. Kaplan propose that this instability is due to chromosomes coming unglued from the molecular tethers that are responsible for ordering their movements.

Green and Kaplan noticed that colorectal tumour cells, known to have high rates of chromosome instability, also showed a loss of microtubules in the spindle apparatus. This is the delicate intracellular network of fibres that coordinate the distribution of chromosomes during cell division. A more detailed look revealed that much of the microtubule loss occurred at the sites where the fibres attach to each chromosome. As a result, the chromosomes did not line up or move appropriately as the cells divided.

Coincidentally, these attachment sites also contain the adenomatous polyposis coli (APC) protein, the gene for which is frequently mutated in human colorectal cancer. When the researchers introduced a mutated form of APC into healthy cells, they saw a similar disruption of the spindle and abnormal chromosome movements to that observed in tumour cells. Thus, a similar mutation in healthy intestinal cells could predispose the cells to further genetic abnormalities and eventually lead to tumour formation.

Brian Fiske



Internet connections: the world's e-mail network.

Secrets of successful stone-skipping

Hitting the water at a magic angle gives top performance in a time-honoured pastime.

Skipping stones across water has been a popular pastime for thousands of years — the rules of the game have remained unchanged since the time of the ancient Greeks¹ — and the world record, set by J. Coleman-McGhee in 1992, is believed to be 38 rebounds². Following earlier attempts^{3–6} to analyse the physics of this ancestral human activity, we focus here on the crucial moment in stone skipping: when the stone bounces on the water's surface. By monitoring the collision of a spinning disc with water, we have discovered that an angle of about 20° between the stone and the water's surface is optimal with respect to the throwing conditions and yields the maximum possible number of bounces.

A stone-skipping throw involves four parameters (Fig. 1): U and Ω are the translational and spin velocities, respectively, α is the 'attack' angle of the stone in relation to the water's surface, and β is the impact angle of the translational velocity. Our experimental set-up is designed to control for each of these parameters independently. Collision sequences were recorded using a high-speed video camera (Fig. 1a), allowing factors such as collision time, change in orientation of the stone, and the shape of the liquid cavity to be determined.

Regarding the role of spin velocity, rotation is found to stabilize the stone (as expected⁴) owing to a gyroscopic effect. We focus our analysis on the high-spin velocity limit, at which α remains constant along the impact. A dynamic phase diagram can then be constructed using the three remaining control parameters (U , α and β), highlighting the necessary conditions for a successful bounce (the 'skipping-stone' domain). Cross-sections in the $\{U, \alpha\}$ and $\{\alpha, \beta\}$ variables are shown in Fig. 1b, c.

The value $\alpha \approx 20^\circ$ is unexpectedly found to play a specific role in this phase diagram: the lowest velocity for a rebound, $U_{\min}$, reaches a minimum for $\alpha \approx 20^\circ$, whereas the maximal successful domain in the impact angle, β , is also achieved for this specific value of α . It can be seen that no rebound is possible for impact angles that are larger than 45°. In a quantitative analysis of the collision, experimental measurements for the collision time, τ , are shown in Fig. 1d: the main feature on this plot is the existence of a minimal value of the collision time, $\tau_{\min}$, which is obtained for $\alpha \approx \alpha_{\min} \approx 20^\circ$ for all velocities.

This minimal collision time is found to obey a simple scaling when the velocity, U , radius, R , and thickness, e , of the stone are varied: namely, $\tau_{\min} \propto \sqrt{(eR)/U}$ (for fixed α

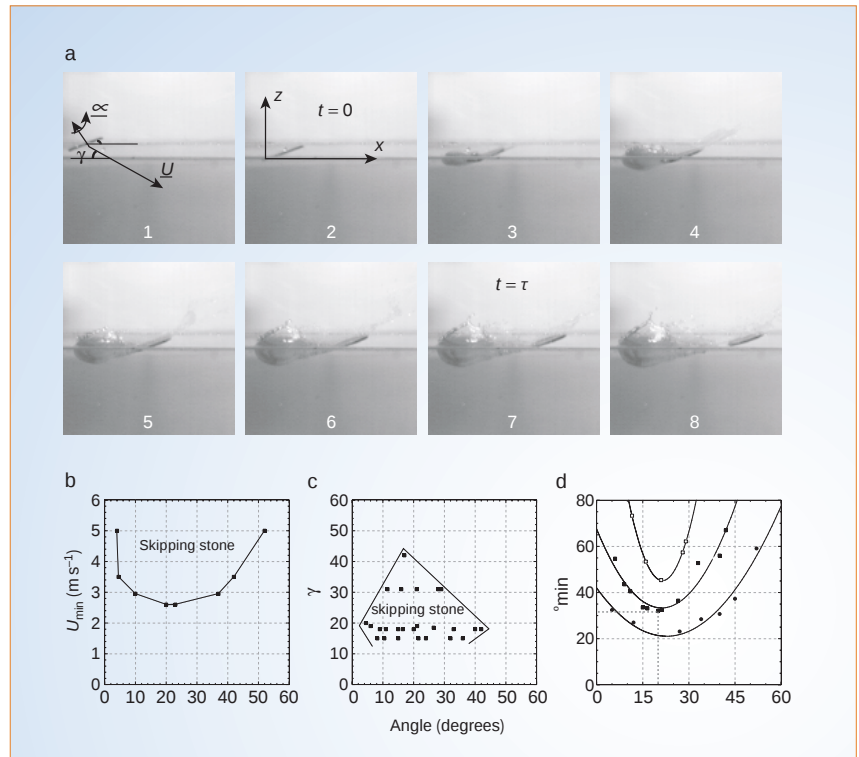


Figure 1 Analysis of stone-skipping. **a**, Chronological photography of a skipping stone, using an aluminium disc as a model stone (radius, $R = 2.5$ cm; thickness, $e = 2.75$ mm; translation velocity, $U = 3.5$ m s^{-1} ; angular velocity, $\Omega = 65$ rotations s^{-1} ; attack angle, $\alpha = 20^\circ$; and trajectory angle, $\beta = 20^\circ$). Time increases from left to right and from top to bottom, with time step $\Delta t = 6.5$ ms. **b–d**, Definition of the skipping-stone domain (for the model stone in **a**): **b**, domain of the skipping stone in the $\{U, \alpha\}$ plane for a fixed $\beta = 20^\circ$; **c**, domain of the skipping stone in the $\{\alpha, \beta\}$ plane for a fixed $U = 3.5$ m s^{-1} . The domain $\beta < 15^\circ$ was not attainable in our experimental set-up. **d**, Evolution of the collision time, τ , as a function of the attack angle, α , under various conditions: filled squares, $U = 3.5$ m s^{-1} , $\beta = 20^\circ$; open squares, $U = 3.5$ m s^{-1} , $\beta = 30^\circ$; circles, $U = 5.0$ m s^{-1} , $\beta = 20^\circ$. The density ratio of the aluminium stone (s) to water (w) was $\rho_s/\rho_w \approx 2.7$ in all experiments.

and β). This scaling is inferred from a simple dimensional analysis. As the lift force, F_{lift} , is the key point in the rebounding process, a collision time can be constructed from the dynamical law as $\tau \propto \sqrt{(mR/F_{\text{lift}})}$, where m is the mass of the stone.

For the velocities under consideration, the lift force is expected to scale as $F_{\text{lift}} \propto \rho_w S_{\text{wetted}} U^2$, where ρ_w is the mass density of water and the wetted area scales as $S_{\text{wetted}} \propto \pi R^2$ (refs 7–9). Using $m = \rho_s e \pi R^2$, where ρ_s is the mass density of the stone, then $\tau \propto \sqrt{(\rho_s/\rho_w)\sqrt{(eR)/U}}$, as measured experimentally. This simple argument, although informative, does not help in understanding the existence of a minimum collision time, a behaviour that requires a more detailed description of the time-dependent hydrodynamic flow around the stone.

The 'magic' angle $\alpha \approx 20^\circ$ is accordingly expected to maximize the number of bounces because the amount of energy dissipated during a collision is directly propor-

tional to the collision time⁴. The ancient art of stone-skipping may therefore benefit from modern scientific insight.

Christophe Clanet*, **Fabien Hersen†**, **Lydéric Bocquet‡**

**Institut de Recherche sur les Phénomènes Hors Équilibre, UMR 6594 du CNRS, BP 146, 13384 Marseille, France*

e-mail: clanet@irphe.univ-mrs.fr

†Ecole Polytechnique, 91128 Palaiseau, France

‡Laboratoire PMCN, UMR CNRS 5586, Université Lyon-I, 69622 Villeurbanne, France

1. D'Arcy Thomson *Alliage* **44**, 77–78 (2000).

2. <http://www.stoneskipping.com>

3. *Encyclopedia Diderot-d'Alembert* 221–222 (Paris, 1744 and 1751).

4. Bocquet, L. *Am. J. Phys.* **71**, 150–155 (2001).

5. Stong, C. L. *Sci. Am.* **219**, 112–118 (1968).

6. Crane, H. R. *Phys. Teach.* **26**, 300–301 (1988).

7. Guyon, E., Hulin, J.-P. & Petit, L. *Hydrodynamique Physique* (CNRS, Paris, 1991).

8. Tritton, D. J. in *Physical Fluid Dynamics* 2nd edn 97–105 (Oxford Univ. Press, Oxford, 1988).

9. Landau, L. D. & Lifshitz, E. M. in *Fluid Mechanics* 168–175 (Pergamon, Oxford, 1959).

Competing financial interests: declared none.

Plant degradation

A nematode expansin acting on plants

Expansin proteins, which have so far been identified only in plants, rapidly induce extension of plant cell walls by weakening the non-covalent interactions that help to maintain their integrity¹. Here we show that an animal, the plant-parasitic roundworm *Globodera rostochiensis*, can also produce a functional expansin, which it uses to loosen cell walls when invading its host plant. As this nematode is known to be able to disrupt covalent bonds in plant cell walls^{2,3}, its accompanying ability to loosen non-covalent bonds challenges the prevailing view that animals are genetically poorly equipped to degrade plant cell walls.

The plant cell wall is a rigid network of interwoven polymers, and many organisms that use plants as a food source use a variety of glycanase enzymes to break covalent bonds in this polysaccharide-based structure. We used complementary DNA–AFLP (for amplified fragment-length polymorphism)-based transcript profiling of synchronized life stages as a starting point to identify the proteins that are used by *G. rostochiensis* to degrade cell walls. A cDNA fragment, KT21 (137 nucleotides), was found to be predominantly expressed in infective second-stage juveniles (J2) and the corresponding full-length cDNA (*Gr-Exp1*; accession number AJ311901; length, 1,061 base pairs) encoded a protein of 271 amino acids that has a predicted amino-terminal signal peptide for secretion.

Similarity searches (BLASTP) indicated that two distinct regions are present in the predicted mature protein. Domain 1 (residues 26–118) shows significant similarity to the carbohydrate-binding module family II of endoglucanases (AF056110, BAB68522 and AF323087; 39–43% identity and expectation values (*E*-value) from 2.0×10^{-12} to 0.0008) from various nematode species. Domain 2 (residues 150–271) showed significant similarity to a β -expansin-like protein (PPAL) from *Nicotiana tabacum* (AAG52887; *E*-value, 2.2×10^{-5}) and a putative β -expansin from *Arabidopsis thaliana* (O04484; *E*-value, 6×10^{-4}). A local alignment of domain 2 with these β -expansins indicated the presence of a series of conserved cysteine residues, the HFD motif (although *Gr-EXP1* harbours a conservative substitution (F→V)) and other conserved motifs⁴.

Whole-mount *in situ* hybridization was carried out on pre-parasitic infective second-stage juveniles of *G. rostochiensis*⁵. Antisense cDNA probes amplified from the *Gr-Exp1* cDNA (nucleotides 54–427) hybridized specifically to the subventral oesophageal glands (Fig. 1a). *Gr-EXP1* antiserum reacted strongly with nematode secretions, induced

by potato-root diffusate, on dot blots (results not shown). We conclude that *Gr-EXP1* is produced in the subventral oesophageal glands of infective juveniles, and that *Gr-EXP1* and cell-wall-degrading enzymes⁶ are secreted simultaneously.

Cell-wall-extension activity⁷ was demonstrated in homogenates of infective second-stage juveniles, and was much stronger on wheat than on cucumber (Fig. 1b). Homogenates of adult females showed no such activity (Fig. 1b). Protein extracts from mature leaves of *Gr-Exp1*-transformed tobacco produced significantly more expansin activity on wheat coleoptiles than did empty-vector controls (Fig. 1c). On the basis of the significant similarity of *Gr-EXP1* to putative β -expansins from *N. tabacum* and *A. thaliana*, the presence of several amino-acid motifs that are characteristic of expansins, and the potent expansin activity

of both recombinant *Gr-EXP1* and nematode homogenates on plant-cell walls, we conclude that *Gr-Exp1* encodes a functional expansin.

Sequences that remotely resemble expansins have been found in various taxa outside the plant kingdom^{8–10}, but cell-wall-loosening activity has not been demonstrated for any of the corresponding proteins. To our knowledge, *Gr-Exp1* is the first non-plant gene found to have the structural and functional characteristics that define the expansin superfamily.

This finding undermines the previously accepted view that animals are poorly equipped for degrading plant cell walls. When cell-wall-degrading enzymes and expansin are simultaneously secreted by the cyst nematode, the activity of expansin may increase the accessibility of cell-wall components to glycanases. This might account for the remarkably high rate (about 2 min per cell layer) at which cyst nematodes can penetrate the host plant.

Ling Qin*, **Ursula Kudla***,
Erwin H. A. Roze*, **Aska Goverse***,
Herman Popeijus†, **Jeroen Nieuwland‡**,
Hein Overmars*, **John T. Jones§**,
Arjen Schots†, **Geert Smant***, **Jaap Bakker***,
Johannes Helder*

*Laboratory of Nematology and †Laboratory of Molecular Recognition and Antibody Technology, Graduate School of Experimental Plant Sciences, Wageningen University, 6709 PD Wageningen, The Netherlands

e-mail: hans.helder@wur.nl

‡Department of Experimental Botany, Graduate School Experimental Plant Sciences, Catholic University of Nijmegen, 6525 ED Nijmegen, The Netherlands

§Plant–Pathogen Interactions Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

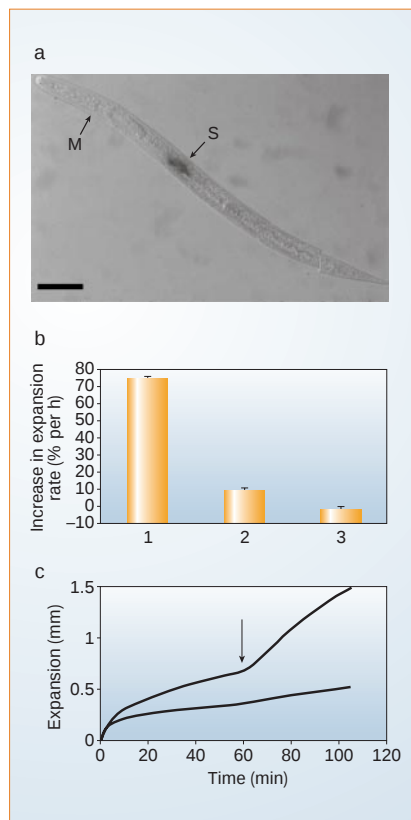


Figure 1 Localization of the nematode *Gr-Exp1* transcript and extension activity of *Gr-EXP1* on plant cell walls. **a**, *In situ* hybridization labelling pattern in infective second-stage juveniles (J2) using a *Gr-Exp1* antisense probe. Arrows indicate the subventral glands (S) and the metacarpus (M), respectively. Scale bar, 20 μ m. **b**, Effects of nematode homogenate on the extension rate of heat-inactivated wheat coleoptiles and cucumber hypocotyls. Bar 1, J2 homogenate on wheat coleoptiles; bar 2, J2 homogenate on cucumber hypocotyls; bar 3, young female homogenate on wheat coleoptiles. **c**, Extension curves for a wheat coleoptile treated with mature-leaf extracts from *Gr-Exp1*-harbouring (upper curve) and empty-vector-harboring (lower curve) tobacco plants. Arrow indicates the point at which the control buffer was replaced by leaf extract.

- Cosgrove, D. J. *Nature* **407**, 321–326 (2000).
- Popeijus, H. et al. *Nature* **406**, 36–37 (2000).
- Jaubert, S., Laffaire, J. B., Abad, P. & Rosso, M. N. *FEBS Lett.* **522**, 109–112 (2002).
- Cosgrove, D. J. *Curr. Opin. Plant Biol.* **3**, 73–78 (2000).
- Qin, L. et al. *Mol. Plant Microbe Interact.* **13**, 830–836 (2000).
- Davis, E. L. et al. *Annu. Rev. Phytopathol.* **38**, 365–396 (2000).
- McQueen, M. S., Durachko, D. M. & Cosgrove, D. J. *Plant Cell* **4**, 1425–1433 (1992).
- Laine, M. J. et al. *Physiol. Mol. Plant Pathol.* **57**, 221–233 (2000).
- Saloheimo, M. et al. *Eur. J. Biochem.* **269**, 4202–4211 (2002).
- Xu, B., Hellman, U., Ersson, B. & Jansson, J. C. *Eur. J. Biochem.* **267**, 4970–4977 (2000).

Competing financial interests: declared none.

brief communications is intended to provide a forum for brief, topical reports of general scientific interest and for technical discussion of recently published material of particular interest to non-specialist readers (**communications arising**). Priority will be given to contributions that have fewer than 500 words, 10 references and only one figure. Detailed guidelines are available on *Nature's* website (www.nature.com/nature).

Reconstructing galaxy histories from globular clusters

Michael J. West¹, Patrick Côté², Ronald O. Marzke³ & Andrés Jordán²

¹Department of Physics and Astronomy, University of Hawaii, 200 W. Kawili Street, Hilo, Hawaii 96720, USA

²Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, New Jersey 08854, USA

³Department of Physics and Astronomy, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132, USA

Nearly a century after the true nature of galaxies as distant ‘island universes’ was established, their origin and evolution remain great unsolved problems of modern astrophysics. One of the most promising ways to investigate galaxy formation is to study the ubiquitous globular star clusters that surround most galaxies. Globular clusters are compact groups of up to a few million stars. They generally formed early in the history of the Universe, but have survived the interactions and mergers that alter substantially their parent galaxies. Recent advances in our understanding of the globular cluster systems of the Milky Way and other galaxies point to a complex picture of galaxy genesis driven by cannibalism, collisions, bursts of star formation and other tumultuous events.

One way to study the history of galaxies is by observing those that are far away. Because light travels at a finite speed, we see distant objects as they looked in the past. By observing galaxies over a wide range of distances, astronomers are able to look back in time to see how galaxies were at different epochs, with the goal of someday constructing a complete chronology of galaxy evolution. In practice, however, the same great distances that make remote galaxies interesting targets for study also make them difficult to observe^{1–3}. Additionally, the inherently statistical nature of this approach, although providing information about the evolution of galaxies as a population over cosmological timescales, can say little about the unique history of any particular galaxy^{4–6}.

An alternative way to learn about galaxy origins is to study those galaxies that are nearby. Because the properties of galaxies today reflect both the conditions at the time of their formation and the cumulative effects of billions of years of evolution, careful analysis of present-day galaxies can yield clues to their past. But the success of this approach hinges on the ability to identify some property of galaxies (or some ‘tracer’ population within them) that can serve as a reliable gauge of their evolution over billions of years. Here we focus on one of the most promising of tracer populations—globular clusters. Remarkable progress has occurred over the past decade in our understanding of the globular cluster systems of galaxies, in particular the discovery that many galaxies possess two or more distinct subpopulations of globular clusters. An underlying theme of this Review is that astronomers are on the verge of reconstructing the individual formation histories of large numbers of galaxies from careful analysis of their globular cluster populations.

Diagnostics of galaxy formation

Globular clusters are dense, gravitationally bound collections of hundreds of thousands to millions of stars that share a common age and chemical composition. Most galaxies are surrounded by systems of tens, hundreds or even thousands of globular clusters that swarm about them like bees around a hive. An example is shown in Fig. 1. In general, the number of globular clusters that a galaxy possesses is roughly proportional to its luminosity⁷.

The earliest studies of globular clusters were limited to those belonging to the Milky Way and its closest galactic neighbours, as the globular cluster systems of more distant galaxies were, with a few notable exceptions^{8–10}, too faint to be seen in significant numbers before the advent of sensitive electronic detectors for astronomical telescopes.

The ages of the Milky Way’s approximately 150 globular clusters, which can be determined by comparing the colours and luminosities of their constituent stars with models of stellar evolution, are found to be in the range 10 to 15 billion years¹¹. This makes them among the most ancient objects in the Universe, fossils from the birth of our Galaxy that have survived to the present. For this reason, globular clusters have played a key role in the continuing debate over how the Milky Way formed. Searle and Zinn’s analysis¹² of the globular cluster population in the halo of the Milky Way, for example, led them to conclude that the halo of our Galaxy formed via the slow accretion of many small proto-galactic fragments, contrary to the then-prevailing view of a more rapid and monolithic collapse¹³. Mounting evidence supports this view of the Milky Way as a ‘cannibal’ that may have consumed dozens or perhaps hundreds of smaller galaxies or galactic fragments during its lifetime^{14–18}. Such cannibalism appears to be quite common among galaxies^{19,20}. Although this process eventually erases the structural identities of progenitor galaxies, their stars and globular clusters remain intact, providing a rich source of historical detail about the birth of present-day galaxies.

Globular clusters are especially attractive probes of galaxy formation for a number of reasons. Foremost is the evidence that some properties of globular cluster systems are correlated with those of their host galaxies^{7,21,22}, which suggests that the birth of these star clusters is linked to the formation of the galaxy itself and hence that they can be used to tease out information about the process of galaxy building. Furthermore, unlike the diffuse stellar population in galaxies, which spans a wide range of ages and chemical abundances, the stars within a given globular cluster are a coeval and extremely homogeneous population, which makes them much simpler stellar systems to understand. Additionally, their luminosity and compact size make them among the brightest and most readily identifiable objects within galaxies, allowing them to be traced to large galactocentric distances where observations of the diffuse stellar component of the galaxy become impractical because of the precipitous drop in surface brightness.

Globular cluster subpopulations

Advances in astronomical imaging and spectroscopic capabilities over the past decade, especially the exquisite image quality provided by the Hubble Space Telescope, have led to tremendous progress in our understanding of the globular cluster systems of galaxies beyond the Milky Way. Two key discoveries in particular have forced a rethinking of the earlier view of globular clusters as a

homogeneous population of ancient fossils left over from the earliest stages of galaxy formation.

First, the detection of young massive star clusters in starburst and merging galaxies whose sizes, luminosities and colours suggest that they may be newly formed globular clusters^{23–27}. The presence of such objects in disturbed galaxies suggests that the disturbances themselves may act as a catalyst for the birth of globular clusters, and raises the question of whether this might be the primary mechanism of cluster formation. Age dating of such clusters can provide constraints on the frequency and duration of galaxy interactions^{28,29}.

This first discovery was soon followed by the second: namely, that many—perhaps most—large galaxies possess two or more subpopulations of globular clusters that have quite different chemical compositions^{30–33}. Figure 2 shows histograms of globular cluster metallicities for several nearby elliptical galaxies. By convention, the metallicity of stars and star clusters, which astronomers denote $[Fe/H]$, is measured relative to the Sun's chemical abundance:

$$\left[\frac{Fe}{H}\right] = \log\left(\frac{N_{Fe}}{N_H}\right) - \log\left(\frac{N_{Fe}}{N_H}\right)_{Sun}$$

where N_{Fe}/N_H is the relative numbers of iron and hydrogen atoms, a good indicator of overall metallicity. Globular cluster metallicities are usually inferred from their photometric colours, which serve as a proxy for the more reliable but also more observationally challenging spectroscopic measurement of chemical abundances^{34–39}. Two peaks are clearly seen in the globular cluster metallicity distribution of each galaxy in Fig. 2, and statistical modelling confirms that these are well described by the sum of two gaussian distributions with distinct metal-poor and metal-rich peaks. Subsequent studies revealed that these different globular cluster subpopulations also have different spatial distributions and kinematics^{40–45}.

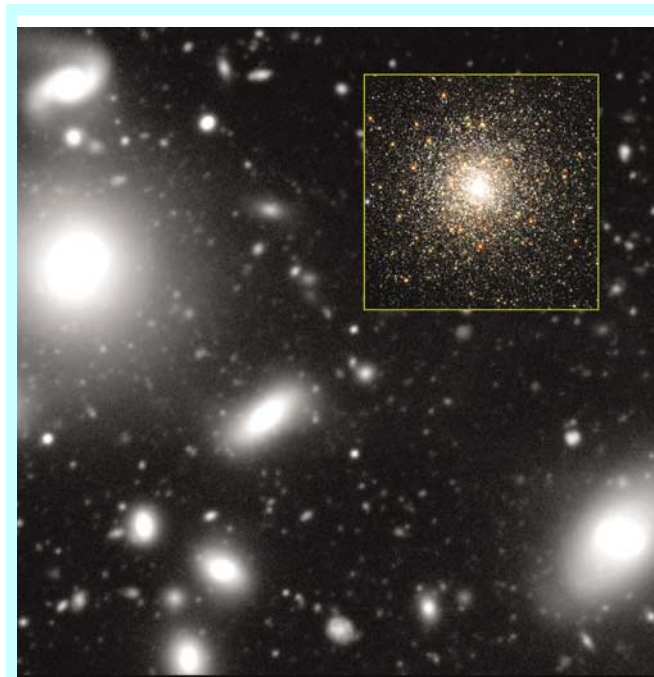


Figure 1 Extragalactic globular cluster systems. Globular clusters are seen as a swarm of faint point-like sources surrounding galaxies in this deep image obtained with the Keck I telescope. Inset, a Hubble Space Telescope image of a nearby globular cluster in our own Milky Way galaxy. A typical globular cluster is composed of as many as a million stars. Most galaxies possess globular cluster systems, from dwarf galaxies with just a handful of clusters to giant galaxies with tens of thousands of clusters. Inset image courtesy of NASA and the Hubble Heritage Team (STScI/AURA).

With hindsight, the ubiquity of multiple globular cluster populations should perhaps not have come as any great surprise, as two distinct globular cluster subsystems—both old—have long been known to co-exist within the Milky Way^{46–49}. The majority of Galactic globular clusters belong to a metal-poor ($[Fe/H] < -0.8$) population that resides in a slowly rotating, spherical halo surrounding our Galaxy. A second, metal-rich ($[Fe/H] > -0.8$) population shares the flattened spatial distribution and rapid rotation characteristic of the bulge and disk components of the Milky Way.

Competing formation models

The presence of multiple globular cluster subpopulations within a single galaxy is undoubtedly an important clue to how such galaxies formed. A number of different hypotheses have been proposed to explain this phenomenon.

One possibility is that the metal-poor and metal-rich populations correspond to different generations of globular clusters. In the simplest cosmological theories of structure formation, galaxies are envisioned as having originated billions of years ago from clouds of metal-poor gas, mostly hydrogen and helium, that collapsed under gravity's pull to form stars and clusters of stars, the most massive of which might today be identified as globular clusters. Over time, supernovae—the explosive deaths of massive stars—enrich any remaining intragalactic gas with the thermonuclear ‘ashes’ from their interiors, such as carbon, oxygen and other heavier atoms. If a second generation of globular clusters forms from this enriched gas, then two separate species of clusters would co-exist within the same host galaxy, one younger and more metal rich than the other.

But how to trigger multiple episodes of globular cluster formation? The merger of gas-rich galaxies is one possible mechanism^{50–53}, as shock waves generated by the collision might compress the gas and lead to the birth of new globular clusters to join any extant population. This idea received strong support with the discovery of newly formed massive star clusters in some interacting galaxies that might be young globular clusters. An example is shown in Fig. 3. Even without collisions as a trigger, it

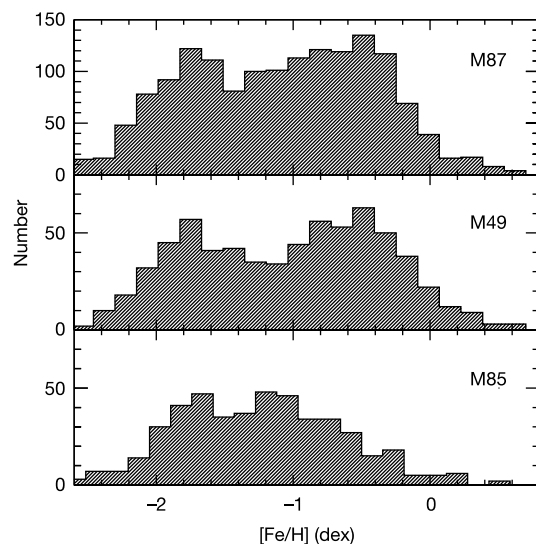


Figure 2 Histograms of globular cluster metallicities in three galaxies. The metallicity, $[Fe/H]$, has been inferred from photometric colours of each galaxy's globular clusters obtained by the authors using the Hubble Space Telescope. Two or more distinct peaks are seen in each case, which indicates the presence of multiple populations of globular clusters with different chemical compositions within these galaxies. A number of different theories have been proposed to explain the origin of these multi-modal globular cluster distributions (see text).

has been suggested that the *in situ* formation of multiple generations of globular clusters might occur under certain conditions within a single galaxy^{54–56}.

A fundamentally different hypothesis for the origin of globular cluster subpopulations in galaxies is to assume that all clusters are roughly coeval and old, and that their numbers are conserved during the hierarchical assembly of galaxies. If, as evidence suggests, large galaxies have grown to their present sizes by cannibalizing smaller ones (see, for example, Fig. 4), or by accreting material stripped from other galaxies during collisions, then the resulting globular cluster systems of such galaxies will be an admixture of their original populations plus those inherited from their victims^{57–61}. Given the preponderance of small galaxies in the Universe, and the observed correlation between galaxy mass and the mean metallicity of its globular clusters^{21,59}, most of the inherited globulars would be more metal-poor than those of the original progenitor galaxy. Galaxy cannibalism and accretion should thus lead to chemically distinct subpopulations of old globular clusters in most large galaxies. Numerical simulations confirm that multi-peaked metallicity distributions like those observed arise quite naturally in such a model without requiring the formation of multiple generations of clusters^{59–61}. Mounting evidence of tidal streams of stars and gas^{15–20,62,63} in the Milky Way and other nearby galaxies—the vestiges of cannibalized companions—lends support to the notion that a significant fraction of their present-day globular cluster populations may have once resided in other galaxies.

Model strengths and weaknesses

Although each of the competing models offers a plausible explanation for the existence of multiple globular cluster populations within the same host galaxy, each also suffers from a number of shortcomings.



Figure 3 Hubble Space Telescope image of a colliding pair of galaxies. The collision appears to have sparked a recent burst of star formation, including the birth of more than 1,000 massive clusters of stars (the prominent blue clumps). The blue colour of these star clusters indicates ages of only a few million years. Some of the most massive objects may evolve over the next few billion years to resemble old globular clusters. This suggests that at least some globular clusters have been born as a result of galaxy collisions and mergers. Image courtesy of NASA/STScI/B. Whitmore.

Models that assume the creation of multiple generations of globular clusters as a result of galaxy collisions or other episodic bursts of star formation are predicated on processes whose physics is poorly understood at present, such as gas heating and cooling, globular cluster formation, the effects of feedback from star formation, and the complex interplay between them. Producing two or more globular cluster subsystems with markedly different metallicities within the same galaxy also requires some fine-tuning. If, for example, globular cluster formation is a prolonged process rather than occurring in bursts at two or more well separated epochs in the metal enrichment histories of galaxies, then the resulting cluster populations would probably have a broad distribution of metallicities lacking the distinct peaks seen in the majority of large galaxies today. Such models are also strongly constrained by the very old ages found for both the metal-rich and metal-poor cluster populations in several of the most thoroughly studied galaxies^{35,39,64–67}. The apparent paucity of young and intermediate aged clusters suggests either that the bulk of galaxy mergers and major star formation episodes must have been completed many billions of years ago, or that such events are not the primary mechanism by which most clusters form.

The accretion model of pre-formed globular clusters also faces a number of challenges. Most obviously, it ignores the creation of new globular clusters that is clearly taking place in some interacting galaxies today, relegating this to a second-order effect rather than the primary formation mechanism for most clusters. Additionally, the accretion model requires a distribution of proto-galactic masses that was very heavily weighted in favour of low-mass objects in order to produce globular cluster metallicity distributions similar to those seen in Fig. 2, with large galaxies having been assembled from hundreds or thousands of cannibalized smaller ones^{59–61}. Such a steep mass function is inconsistent with observations of the luminosity function of galaxies today^{68–71}, but is suggestively similar to the primordial mass spectrum predicted by the popular cold-dark-matter models of structure formation.

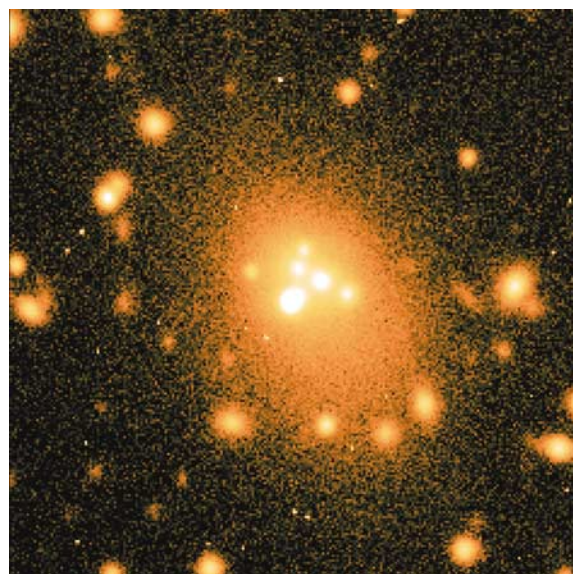


Figure 4 A giant elliptical galaxy in the cluster Abell 3827. The partially digested remains of several smaller cannibalized galaxies are visible in its central region. Globular clusters belonging to those galaxies are likely to survive the eventual disruption of their parent galaxies, and thus will become part of this giant galaxy. If, as evidence suggests, many large elliptical galaxies have grown to their present sizes by cannibalizing smaller neighbours, then their globular cluster populations today are composite systems that may provide clues about the progenitor galaxies.

There is, of course, the possibility that the present-day globular cluster populations of galaxies might have arisen from some combination of collisions, cannibalism and accretion, or perhaps from other processes. For example, a number of authors have suggested that the stripped nuclei of compact dwarf galaxies might masquerade as globular clusters^{72–74}. A good nearby example is the Milky Way's largest globular cluster, ω Centauri, which many astronomers now suspect may be the surviving nucleus of a dwarf galaxy that was gravitationally shredded by our galaxy^{75,76}. Additional complications come from recent discoveries of a new class of faint 'fuzzy' star clusters⁷⁷, 'ultra compact dwarf' galaxies⁷⁸ and a population of intergalactic globular clusters of unknown origin^{79,80}. Although these various objects probably comprise only a small fraction of the globular clusters in normal galaxies, their existence demonstrates that the processes leading to the formation of present-day globular cluster systems are likely to have been many and varied. The key question is what was the dominant process that governed the assembly of present-day globular cluster systems. One conclusion shared by the various competing models is that the genesis of present-day galaxies and their globular cluster systems has been a long, drawn-out process that is still continuing, rather than a single event in the distant past.

Where do we go from here?

Much work remains to be done to better understand extragalactic globular cluster systems and their use as probes of galaxy formation. Significant progress in the near future is likely to occur along several fronts.

On the theoretical front, each of the competing models for the origin of multiple globular cluster populations in galaxies is undoubtedly oversimplified. New insights may come from high-resolution computer simulations or semi-analytic models of galaxy formation that start from more realistic cosmological initial conditions. By following the detailed merger histories of simulated galaxies from early epochs to the present, and including the effects of gas dynamics, star formation, chemical enrichment, and energy input from supernovae, it should be possible to refine quantitative predictions about the evolution of globular cluster systems as a function of time. Important first steps in that direction have already been taken^{81–86}, and within the next few years it is hoped that numerical simulations will achieve sufficient mass resolution to allow the growth of globular clusters, galaxies and larger-scale structures to be followed simultaneously.

On the observational side, although the globular cluster systems of distant galaxies will probably remain inaccessible to direct observation for the foreseeable future⁸⁷, many thousands of nearby galaxies within the grasp of current technology await study. Thanks to the new generation of large, sensitive CCD cameras, including the new Advanced Camera for Surveys on board the Hubble Space Telescope, it is feasible to obtain deep, multi-colour images of large numbers of galaxies with a modest investment of telescope time. A comprehensive census of the globular cluster systems of galaxies in the local Universe would be invaluable, as it would provide a more thorough assessment of the frequency of globular cluster subpopulations, their properties, and variations with galaxy type and environment. An example of what can be done is the recent ACS Virgo Cluster Survey (<http://www.physics.rutgers.edu/pcote/acs>), which will provide data for many thousands of globular clusters associated with 100 early-type galaxies in the nearby Virgo cluster of galaxies. Galaxies that do not have multiple globular cluster populations might also provide insights into the galaxy formation process. Direct comparison of the metallicities of globular clusters and halo stars in the same galaxy, although observationally challenging, is likely to shed additional light on the origin of galaxies^{88,89}.

Major spectroscopic programmes are also needed to obtain large numbers of radial velocities and abundance measurements of globular clusters from high signal-to-noise spectra; such data

would allow more detailed studies to be undertaken of the kinematics, compositions and star formation histories of cluster subpopulations. Although photometric observations can be, and commonly are, used to infer the metallicities of globular clusters, they are no substitute for spectroscopy. Spectroscopic measures of metallicity are much more precise than the photometric measures described earlier, and our ability to produce precise timelines is tied directly to the accuracy of the metallicities. However, the faintness of all but the nearest clusters makes such observations challenging, even with telescopes of the 8–10 m class. The next generation of very large 30–50 m telescopes that are planned for the coming decade should allow great progress to be made in this area.

An especially fertile area of research in the next few years may be age determinations of globular cluster subpopulations in galaxies, as here the competing models make rather different predictions. If the metal-rich and metal-poor clusters are the result of two separate bursts of cluster formation, then the metal-rich globular clusters must be younger than those that are metal-poor. If, on the other hand, the multiple globular cluster populations seen in galaxies can be accounted for by the accretion of pre-existing globulars, then the metal-rich and metal-poor globulars should be predominantly old and coeval, with no obvious correlation between metallicity and age. Whereas current methods of age determinations are not sufficiently precise to rule out the competing models^{35,39,64–67}, improved methods hold great promise for reducing uncertainties to the point where stronger constraints on models can be made. Detailed spectroscopic and photometric observations of the most massive galaxies in dense environments show that their globular cluster populations are generally old and coeval to within a couple of billion years^{35,39,64–67}. But the age distribution of globular clusters remains an open question, as several recent studies have found a wider spread of globular cluster ages in some galaxies, including the presence of some intermediate-age metal-rich clusters^{90–92}. The findings to date suggest a complex picture of globular cluster formation in galaxies, with the formation histories of no two galaxies being exactly alike.

These are exciting times in the study of extragalactic globular cluster systems. Disentangling the myriad processes that may have contributed to the formation of nearby galaxies is a formidable challenge, but the potential reward is spectacular: a glimpse into the evolutionary histories of individual galaxies. □

doi:10.1038/nature02235.

1. Barger, A. J. *et al.* Submillimetre-wavelength detection of dusty star-forming galaxies at high redshift. *Nature* **394**, 248–251 (1998).
2. Blain, A. W., Smail, I., Ivison, R. J. & Kneib, J.-P. The history of star formation in dusty galaxies. *Mon. Not. R. Astron. Soc.* **392**, 632–648 (1999).
3. Abraham, R. G. & van den Bergh, S. The morphological evolution of galaxies. *Science* **293**, 1273–1278 (2001).
4. Cowie, L. L., Songaila, A., Hu, E. & Cohen, J. G. New insight on galaxy formation and evolution from Keck spectroscopy of the Hawaii Deep Fields. *Astron. J.* **112**, 839–864 (1996).
5. Lilly, S. J., Le Fevre, O., Hammer, F. & Crampton, D. The Canada-France Redshift Survey: The luminosity density and star formation history of the universe to z approximately 1. *Astrophys. J.* **460**, L1–L4 (1996).
6. Madau, P. *et al.* High-redshift galaxies in the Hubble Deep Field: Colour selection and star formation history to $z \sim 4$. *Mon. Not. R. Astron. Soc.* **283**, 1388–1404 (1996).
7. Harris, W. E. Globular cluster systems in galaxies beyond the Local Group. *Annu. Rev. Astron. Astrophys.* **29**, 543–579 (1991).
8. Baum, W. A. The distribution of luminosity in elliptical galaxies. *Publ. Astron. Soc. Pacif.* **67**, 328–329 (1955).
9. Racine, R. 2000 globular clusters in M87. *J. R. Astron. Soc. Can.* **62**, 367–376 (2000).
10. Strom, S. E. *et al.* The halo globular clusters of the giant elliptical galaxy Messier 87. *Astrophys. J.* **245**, 416–453 (1981).
11. Krauss, L. M. & Chaboyer, B. Age estimates of globular clusters in the Milky Way: Constraints on cosmology. *Science* **299**, 65–70 (2003).
12. Searle, L. & Zinn, R. Compositions of halo clusters and the formation of the Galactic halo. *Astrophys. J.* **225**, 357–379 (1978).
13. Eggen, O. J., Lynden-Bell, D. & Sandage, A. R. Evidence from the motions of old stars that the Galaxy collapsed. *Astrophys. J.* **136**, 748–766 (1962).
14. Ibata, R. A., Gilmore, G. & Irwin, M. J. A dwarf satellite galaxy in Sagittarius. *Nature* **370**, 194–196 (1994).
15. Majewski, S. R., Munn, J. A. & Hawley, S. L. Absolute proper motions to B approximately 22.5: Evidence for kinematical substructure in halo field stars. *Astrophys. J.* **427**, L37–L41 (1994).

16. Helmi, A., White, S. D. M., de Zeeuw, P. T. & Zhao, H. Debris streams in the solar neighbourhood as relicts from the formation of the Milky Way. *Nature* **402**, 53–55 (1999).
17. Freeman, K. & Bland-Hawthorn, J. The new Galaxy: Signatures of its formation. *Annu. Rev. Astron. Astrophys.* **40**, 487–537 (2002).
18. Newberg, H. *et al.* The ghost of Sagittarius and lumps in the halo of the Milky Way. *Astrophys. J.* **569**, 245–274 (2002).
19. Ferguson, A. M., Irwin, M. J., Ibata, R. A., Lewis, G. F. & Tanvir, N. R. Evidence for stellar substructure in the halo and outer disk of M31. *Astron. J.* **124**, 1452–1463 (2002).
20. Peng, E. W., Ford, H. C., Freeman, K. C. & White, R. L. A young blue tidal stream in NGC 5128. *Astron. J.* **124**, 3144–3156 (2002).
21. Brodie, J. P. & Huchra, J. P. Extragalactic globular clusters. III—Metallicity comparisons and anomalies. *Astrophys. J.* **379**, 157–167 (1991).
22. Santiago, B. X. & Djorgovski, S. Multivariate analysis of globular cluster systems in early-type galaxies. *Mon. Not. R. Astron. Soc.* **261**, 753–765 (1993).
23. Lutz, D. NGC 3597—Formation of an elliptical via merging? *Astron. Astrophys.* **245**, 31–40 (1991).
24. Holtzman, J. A. *et al.* Planetary Camera observations of NGC 1275—Discovery of a central population of compact massive blue star clusters. *Astron. J.* **103**, 691–702 (1992).
25. Whitmore, B. C., Schweizer, F., Leitherer, C., Borne, K. & Robert, C. Hubble Space Telescope discovery of candidate young globular clusters in the merger remnant NGC 7252. *Astron. J.* **106**, 1354–1370 (1993).
26. Whitmore, B. C. & Schweizer, F. Hubble Space Telescope observations of young star clusters in NGC-4038/4039, ‘the antennae’ galaxies. *Astron. J.* **109**, 960–980 (1995).
27. Schweizer, F., Miller, B. W., Whitmore, B. C. & Fall, S. M. Hubble Space Telescope observations of candidate young globular clusters and stellar associations in the recent merger remnant NGC 3921. *Astron. J.* **112**, 1839–1862 (1996).
28. de Grijs, R., O’Connell, R. W. & Gallagher, J. S. The fossil starburst in M82. *Astron. J.* **121**, 768–792 (2001).
29. Ma, J. *et al.* SED, metallicity and age of halo globular clusters in M33. *Astron. Astrophys.* **385**, 404–411 (2002).
30. Gebhardt, K. & Kissler-Patig, M. Globular cluster systems. I. V–I color distributions. *Astron. J.* **118**, 1526–1541 (1999).
31. Forbes, D. A. & Forte, J. C. The connection between globular cluster systems and the host galaxies. *Mon. Not. R. Astron. Soc.* **322**, 257–261 (2001).
32. Kundu, A. & Whitmore, B. C. New insights from HST studies of globular cluster systems. I. Colors, distances, and specific frequencies of 28 elliptical galaxies. *Astron. J.* **121**, 2950–2973 (2001).
33. Larsen, S. S., Brodie, J. P., Huchra, J. P., Forbes, D. A. & Grillmair, C. J. Properties of globular cluster systems in nearby early-type galaxies. *Astron. J.* **121**, 2974–2998 (2001).
34. Couture, J., Harris, W. E. & Allwright, J. W. B. VBI photometry of globular clusters in M87. *Astrophys. J. Suppl.* **73**, 671–683 (1990).
35. Cohen, J. G., Blakeslee, J. P. & Ryzhov, A. The ages and abundances of a large sample of M87 globular clusters. *Astrophys. J.* **496**, 808–826 (1998).
36. Kissler-Patig, M. *et al.* Keck spectroscopy of globular clusters around NGC 1399. *Astron. J.* **115**, 105–120 (1998).
37. Barmby, P. *et al.* M31 globular clusters: Colors and metallicities. *Astron. J.* **119**, 727–747 (2000).
38. Schulz, J., Fritze-v. Alvensleben, U., Möller, C. S. & Fricke, K. J. Spectral and photometric evolution of simple stellar populations at various metallicities. *Astron. Astrophys.* **392**, 1–11 (2002).
39. Cohen, J. G., Blakeslee, J. P. & Côté, P. The ages and abundances of a sample of globular clusters in M49 (NGC 4472). *Astrophys. J.* **592**, 866–883 (2003).
40. Ashman, K. M. & Bird, C. M. Globular cluster clustering in M31. *Astron. J.* **106**, 2281–2290 (1993).
41. Zepf, S. E. *et al.* Dynamical constraints on the formation of NGC 4472 and its globular clusters. *Astron. J.* **120**, 2928–2937 (2000).
42. Côté, P. *et al.* Dynamics of the globular cluster system associated with M87 (NGC 4496). II. Analysis. *Astrophys. J.* **559**, 828–850 (2001).
43. Schroeder, L. L., Brodie, J. P., Kissler-Patig, M., Huchra, J. P. & Phillips, A. C. Spectroscopy of globular clusters in M81. *Astron. J.* **123**, 2473–2489 (2002).
44. Côté, P., McLaughlin, D. E., Cohen, J. G. & Blakeslee, J. P. Dynamics of the globular cluster system associated with M49 (NGC 4472): Cluster orbital properties and the distribution of dark matter. *Astrophys. J.* **591**, 850–877 (2003).
45. Perrett, K. M., Stiff, D. A. & Hanes, D. A. Substructure in the Andromeda galaxy globular cluster system. *Astrophys. J.* **589**, 790–797 (2003).
46. Kinman, T. D. Globular clusters. II. The spectral types of individual stars and of the integrated light. *Mon. Not. R. Astron. Soc.* **119**, 538–558 (1959).
47. Zinn, R. The globular cluster system of the galaxy. IV—The halo and disk subsystems. *Astrophys. J.* **293**, 424–444 (1985).
48. Minniti, D. Metal-rich globular clusters with R less than or equal 3 kpc: Disk or bulge clusters. *Astron. J.* **109**, 1663–1669 (1995).
49. Côté, P. Kinematics of the Galactic globular cluster system: New radial velocities for clusters in the direction of the inner Galaxy. *Astron. J.* **118**, 406–420 (1999).
50. van den Bergh, S. Star clusters in IC 1613 and in the Small Magellanic Cloud. *Astrophys. J.* **230**, 95–96 (1979).
51. Schweizer, F. in *Nearly Normal Galaxies: From the Planck Time to the Present* (ed. Faber, S.) 18–25 (Proc. 8th Santa Cruz Summer Workshop in Astron. and Astrophys., Springer, New York, 1987).
52. Ashman, K. M. & Zepf, S. E. The formation of globular clusters in merging and interacting galaxies. *Astrophys. J.* **384**, 50–61 (1992).
53. Zepf, S. E. & Ashman, K. M. Globular cluster systems formed in galaxy mergers. *Mon. Not. R. Astron. Soc.* **264**, 611–618 (1993).
54. Forbes, D. A., Brodie, J. P. & Grillmair, C. J. On the origin of globular clusters in elliptical and cD Galaxies. *Astron. J.* **113**, 1652–1665 (1997).
55. Harris, W. E., Harris, G. L. H. & McLaughlin, D. E. M87, globular clusters, and galactic winds: Issues in giant galaxy formation. *Astron. J.* **115**, 1801–1822 (1998).
56. van den Bergh, S. How did globular clusters form? *Astrophys. J.* **559**, L113–L114 (2001).
57. Forte, J. C., Martínez, R. E. & Muzzio, J. C. On the origin of the globular cluster system of M87. *Astron. J.* **87**, 1465–1469 (1982).
58. Muzzio, J. C., Martínez, R. E. & Rabolli, M. Globular cluster swapping in clusters of galaxies. *Astrophys. J.* **285**, 7–15 (1984).
59. Côté, P., Marzke, R. O. & West, M. J. The formation of giant elliptical galaxies and their globular cluster systems. *Astrophys. J.* **501**, 554–570 (1998).
60. Côté, P., Marzke, R. O., West, M. J. & Minniti, D. Evidence for the hierarchical formation of the Galactic spheroid. *Astrophys. J.* **533**, 869–883 (2000).
61. Côté, P., West, M. J. & Marzke, R. O. Globular cluster systems and the missing satellite problem: Implications for Cold Dark Matter models. *Astrophys. J.* **567**, 853–864 (2002).
62. Ibata, R. A., Irwin, M. J., Lewis, G. F., Ferguson, A. M. N. & Tanvir, N. One ring to encompass them all: A giant stellar structure that surrounds the Galaxy. *Mon. Not. R. Astron. Soc.* **340**, L21–L27 (2003).
63. Morrison, H. L., Harding, P., Hurley-Keller, D. & Jacoby, G. Andromeda VIII: A new tidally distorted satellite of M31. *Astrophys. J.* **596**, L183–L188 (2003).
64. Puzia, T. H., Kissler-Patig, M., Brodie, J. P. & Huchra, J. P. The age difference between the globular cluster subpopulations in NGC 4472. *Astron. J.* **118**, 2734–2750 (1999).
65. Forbes, D. A., Beasley, M. A., Brodie, J. P. & Kissler-Patig, M. Age estimates for globular clusters in NGC 1399. *Astrophys. J.* **563**, L143–L146 (2001).
66. Jordán, A., Côté, P., West, M. J. & Marzke, R. O. The relative ages of the globular cluster subpopulations in M87. *Astrophys. J.* **576**, L113–L116 (2002).
67. Larsen, S. & Brodie, J. P. Keck spectroscopy of young and old star clusters in NGC 1023. *Astrophys. J.* **123**, 1488–1496 (2002).
68. Schechter, P. An analytic expression for the luminosity function for galaxies. *Astrophys. J.* **203**, 297–306 (1976).
69. Marzke, R. O., Huchra, J. P. & Geller, M. J. The luminosity function of the CfA Redshift Survey. *Astrophys. J.* **428**, 43–50 (1994).
70. Klypin, A., Kravtsov, A. V., Valenzuela, O. & Prada, F. Where are the missing galactic satellites? *Astrophys. J.* **522**, 82–92 (1999).
71. Moore, B. *et al.* Dark matter substructure within galactic halos. *Astrophys. J.* **524**, L19–L22 (1999).
72. Zinnecker, H., Keable, C. J., Dunlop, J. S., Cannon, R. D. & Griffiths, W. K. In *The Harlow-Shapley Symposium on Globular Cluster Systems in Galaxies* (eds Grindlay, J. E. & Davis Philip, A. G.) 603–604 (Proc. 126th IAU Symp., Kluwer, Dordrecht, 1988).
73. Bassino, L. P., Muzzio, J. C. & Rabolli, M. Are globular clusters the nuclei of cannibalized dwarf galaxies? *Astrophys. J.* **431**, 634–639 (1994).
74. Mieske, S., Hilker, M. & Infante, L. Ultra compact objects in the Fornax cluster of galaxies: Globular clusters or dwarf galaxies? *Astron. Astrophys.* **383**, 823–837 (2002).
75. Gnedin, O. Y. *et al.* The unique history of the globular cluster Omega Centauri. *Astrophys. J.* **568**, L23–L26 (2002).
76. Tsuchiya, T., Dinescu, D. I. & Korchagin, V. A capture scenario for the globular cluster Omega Centauri. *Astrophys. J.* **589**, L29–L32 (2003).
77. Brodie, J. P. & Larsen, S. S. New members of the cluster family in nearby lenticular galaxies. *Astron. J.* **124**, 1410–1417 (2002).
78. Drinkwater, M. J. *et al.* A class of compact dwarf galaxies from disruptive processes in galaxy clusters. *Nature* **423**, 519–521 (2003).
79. West, M. J., Côté, P., Jones, C., Forman, W. & Marzke, R. O. Intracluster globular clusters. *Astrophys. J.* **453**, L77–L80 (1995).
80. Jordán, A., West, M. J., Côté, P. & Marzke, R. O. A point-source excess in Abell 1185: Intergalactic globular clusters? *Astron. J.* **125**, 1642–1648 (2003).
81. Abel, T., Anninos, P., Norman, M. L. & Zhang, Y. First structure formation. I. Primordial star-forming regions in hierarchical models. *Astrophys. J.* **508**, 518–609 (1998).
82. Weil, M. L. & Pudritz, R. E. Cosmological evolution of supergiant star-forming clouds. *Astrophys. J.* **556**, 164–176 (2001).
83. Beasley, M. A., Baugh, C. M., Forbes, D. A., Sharples, R. A. & Frenk, C. S. On the formation of globular cluster systems in a hierarchical Universe. *Mon. Not. R. Astron. Soc.* **333**, 383–399 (2002).
84. Bekki, K. & Chiba, M. Formation of Galactic disk globular clusters in early dissipative minor merging. *Astrophys. J.* **566**, 245–251 (2002).
85. Bromm, V. & Clarke, C. J. The formation of the first globular clusters in dwarf galaxies before the epoch of reionization. *Astrophys. J.* **566**, L1–L4 (2002).
86. Kravtsov, A. V. & Gnedin, O. Y. Formation of globular clusters in hierarchical cosmology. *Astrophys. J.* (in the press).
87. Carlberg, R. G. Globular clusters at high redshift. *Astrophys. J.* **573**, 60–65 (2002).
88. Harris, W. E. & Harris, G. L. H. On the formation of galaxy halos: Comparing NGC 5128 and the Local Group members. *Astron. J.* **122**, 3065–3069 (2001).
89. Beasley, M. A., Harris, W. E., Harris, G. L. H. & Forbes, D. A. Simultaneous modeling of the stellar halo and globular cluster system of NGC 5128. *Mon. Not. R. Astron. Soc.* **340**, 341–352 (2003).
90. Goudfrooij, P., Mack, J., Kissler-Patig, M., Meylan, G. & Minniti, D. Kinematics, ages and metallicities of star clusters in NGC 1316: A 3-Gyr-old merger remnant. *Mon. Not. R. Astron. Soc.* **322**, 643–657 (2001).
91. Jiang, L. *et al.* Spectral energy distributions and age estimates of 172 globular clusters in M31. *Astron. J.* **125**, 727–741 (2003).
92. Larsen, S. S. *et al.* Evidence for an intermediate-age, metal-rich population of globular clusters in NGC 4365. *Astrophys. J.* **585**, 767–774 (2003).

Acknowledgements We thank J. Brodie, W. Harris and S. van den Bergh for comments that helped to improve this manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.W. (westm@hawaii.edu)

X-ray structure of a protein-conducting channel

Bert van den Berg^{1*}, William M. Clemons Jr^{1*}, Ian Collinson², Yorgo Modis³, Enno Hartmann⁴, Stephen C. Harrison³ & Tom A. Rapoport¹

¹Howard Hughes Medical Institute and Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

²Max Planck Institute of Biophysics, Marie-Curie-Strasse 13-15, D-60439 Frankfurt am Main, Germany

³Howard Hughes Medical Institute, Children's Hospital and Harvard Medical School, 320 Longwood Avenue, Boston, Massachusetts 02115, USA

⁴University Luebeck, Institute for Biology, Ratzeburger Allee 160, Luebeck, D-23538, Germany

* These authors contributed equally to this work

A conserved heterotrimeric membrane protein complex, the Sec61 or SecY complex, forms a protein-conducting channel, allowing polypeptides to be transferred across or integrated into membranes. We report the crystal structure of the complex from *Methanococcus jannaschii* at a resolution of 3.2 Å. The structure suggests that one copy of the heterotrimer serves as a functional translocation channel. The α -subunit has two linked halves, transmembrane segments 1–5 and 6–10, clamped together by the γ -subunit. A cytoplasmic funnel leading into the channel is plugged by a short helix. Plug displacement can open the channel into an 'hourglass' with a ring of hydrophobic residues at its constriction. This ring may form a seal around the translocating polypeptide, hindering the permeation of other molecules. The structure also suggests mechanisms for signal-sequence recognition and for the lateral exit of transmembrane segments of nascent membrane proteins into lipid, and indicates binding sites for partners that provide the driving force for translocation.

A decisive step in the biosynthesis of many secretory and plasma-membrane proteins is their transport across the endoplasmic reticulum (ER) membrane in eukaryotes or across the cytoplasmic membrane in prokaryotes (for a review, see ref. 1). These polypeptides are first targeted to the membrane by hydrophobic amino-acid sequences, which are either cleavable signal sequences or transmembrane segments (TM) of membrane proteins. Soluble proteins, such as those destined for secretion, are subsequently transported across the membrane through a protein-conducting channel with a hydrophilic interior^{2,3}. In the case of membrane proteins, when a hydrophobic TM arrives in the channel, it is released through an opening in the channel wall into the surrounding lipid phase. The capacity of the channel to open laterally towards the lipid and the wide variety of substrates that it must transport distinguish it from many other channels.

An evolutionarily conserved heterotrimeric complex of membrane proteins, called the Sec61 complex in eukaryotes and the SecY complex in eubacteria and archaea, forms the channel (for a review, see ref. 4). The α -subunits (Sec61 α in mammals, Sec61p in *Saccharomyces cerevisiae*, SecY in bacteria and archaea) and γ -subunits (Sec61 γ in mammals, Ssl1p in *S. cerevisiae*, SecE in bacteria and archaea) show significant sequence conservation (see Supplementary Fig. S1). Both subunits are required for cell viability in *S. cerevisiae* and *Escherichia coli*. The β -subunits (Sec61 β in mammals, Sbh in *S. cerevisiae*, SecZ in archaea) are not essential for cell viability in these organisms; they are similar in eukaryotes and archaea, but show no obvious homology to the corresponding SecG subunits in bacteria. The α -subunit forms the channel pore, and it is the crosslinking partner of polypeptide chains passing through the membrane⁵. Reconstitution experiments have shown that the Sec61/SecY complex is the essential membrane component for protein translocation^{6,7}.

The channel itself is a passive conduit for polypeptides and must therefore associate with other components that provide a driving force¹. In co-translational translocation, the major partner is the ribosome. The elongating polypeptide chain moves directly from the ribosome into the associated membrane channel. The energy for translocation comes from GTP hydrolysis during translation. Many (or perhaps all) cells also have post-translational translocation, in

which polypeptides are completed in the cytosol and then transported across the membrane. In yeast (and probably in all eukaryotes), the post-translational translocation partners are another membrane protein complex (the tetrameric Sec62/63p complex) and the luminal protein BiP, a member of the Hsp70 family of ATPases^{8,9}. BiP promotes translocation by acting as a molecular ratchet, preventing the polypeptide chain from sliding back into the cytosol¹⁰. In the eubacterial post-translational pathway, the cytosolic ATPase SecA pushes polypeptides through the channel¹¹. In addition, an electrochemical gradient across the membrane stimulates translocation *in vitro* and is essential *in vivo*¹². Archaea lack SecA and the Sec62/63p complex, and it is unclear how they perform post-translational translocation¹³. Despite the differences between the pathways, most mechanistic aspects of translocation that relate to the channel itself are probably similar. Specifically, in all cases the channel partner—either the ribosome, the Sec62/63p complex or SecA—binds first, and the signal sequence or TM of a translocation substrate associates with the channel subsequently, priming it for polypeptide translocation.

An understanding of the mechanisms that underlie protein translocation requires detailed structural information. Low-resolution structures have been obtained by single-particle electron microscopy (EM) of either the isolated Sec61/SecY complex or the Sec61 complex bound to the ribosome^{14–18}. A recent structure of the *E. coli* SecY complex, derived from electron cryo-microscopy of two-dimensional (2D) crystals in a phospholipid bilayer, indicated the expected number of TM helices, but the resolution (about 8 Å in-plane) was insufficient for their identification¹⁹. We now report the structure of the SecY complex from the archae *M. jannaschii*, determined by X-ray diffraction at 3.2 Å resolution.

Structure determination of the SecY complex

We purified and crystallized the complex from *M. jannaschii* in the detergent diheptanoylphosphatidyl choline. Crystals of selenomethionine-derivatized protein diffract to 3.4 Å and have different unit-cell dimensions, depending on buffer conditions (see Supplementary Table 1). Initial phases were obtained from single-wavelength anomalous dispersion (SAD) experiments and improved by a combination of cross-crystal averaging, solvent

flattening and histogram matching. After obtaining an initial model, we generated point mutants in the α -subunit aimed at strengthening crystal contacts. One double mutant (Lys422Arg and Val423Thr), called Y1, gave crystals that diffracted to 3.2 Å. There are subtle differences between the mutant and wild-type structures (see Supplementary Fig. S2), and we have used the latter to generate the figures below. The final wild-type model includes all residues, with the exception of some at the termini (α -subunit 434–436; β -subunit 1–20, 53; γ -subunit 67–73). Examples of the fit into the experimental electron density map are shown in Supplementary Fig. S3, and a summary of the quality of the structures is given in Table 1.

Architecture of the SecY complex

General characteristics

The crystal structure contains a single copy of the SecY complex with one polypeptide chain of each of the three subunits. Figure 1a shows a stereo view from the cytoplasm ('top view') and Fig. 1b shows the complex from the side (for more details, see Supplementary Fig. S4). As expected (see ref. 4), the α -subunit contains ten TMs with amino and carboxy termini in the cytosol, and the β - and γ -subunits have one TM each, with their N termini in the cytosol. Viewed from the top, the SecY complex has an approximately rectangular shape (Fig. 1a). The α -subunit is open on one side (the 'front') and is surrounded on the remaining three sides by the two smaller subunits. The side view shows that only the cytoplasmic domains protrude significantly beyond the phospholipid-head-group region of the membrane (Fig. 1b).

The α -subunit

This subunit is divided into two halves, TM1–5 and TM6–10 (Fig. 1c), which are connected at the back of the molecule by an external loop between TM5 and TM6. Each half consists of three outer and two inner TMs (Fig. 1d) related by pseudo-symmetry through a two-fold rotation axis in the plane of the membrane (Fig. 1d), and they have similar folds (Supplementary Fig. S5). The second half is essentially an inverted version of the first half, and pairs of helices in the two parts are topologically related (for example, TM2 and TM7). A similar pseudo-symmetry that is not obvious from the primary sequence has also been observed in the structures of several other membrane proteins (for example, refs 20, 21).

Many of the TMs in the α -subunit are not perpendicular to the plane of the membrane (for example, TM2, TM5 and TM7), and

some do not span the entire membrane (for example, TM9 and TM10). In addition, several loops have a complex secondary structure, and some are tucked back into the membrane (for example, the loop between TM7 and TM8). The most striking feature of the α -subunit is the segment following TM1 (Fig. 1e). In most organisms, it begins with the sequence PΦXG (Φ is a conserved hydrophobic amino acid). It then continues as a long loop that runs along the external side of the molecule before leading back into the centre of the α -subunit and ending in a short, distorted helix, called TM2a. This helix extends to a point about halfway through the membrane (Fig. 1e). It is not particularly hydrophobic and was predicted to be in the external aqueous phase²². TM2a is followed by a segment resembling a β -hairpin loop centred on Gly 68 (Supplementary Fig. S3a). This loop ends in the middle of the membrane at the completely conserved Gly 74 of the sequence GΦXP. A sharp turn leads into TM2b, which extends from the centre of the molecule outwards at a $\sim 30^\circ$ angle with respect to the plane of the membrane.

The β -subunit

This polypeptide begins with a disordered cytoplasmic segment followed by a loop that crosses over the N terminus of the α -subunit (Fig. 1a). The TM is almost perpendicular to the plane of the membrane and comes close to the C terminus of the γ -subunit at the external side of the membrane (Fig. 1b). The β -subunit makes only limited contact with the α -subunit, which may explain why it is not essential for the function of the complex.

The γ -subunit

This subunit consists of two helices. The N-terminal helix lies on the cytoplasmic surface of the membrane (Fig. 1a, b). In agreement with predictions and crosslinking studies^{23,24}, this helix is amphipathic with the hydrophobic surface pointing towards the membrane, contacting the C-terminal part of the α -subunit. The helix is followed by a short β -strand that forms a sheet with a segment of the β -hairpin between TM6 and TM7 of the α -subunit (Fig. 1a). The TM of the γ -subunit is a long, curved helix that crosses the membrane at a $\sim 35^\circ$ angle with respect to the plane of the membrane (Fig. 1b). One side of the helix makes limited contacts with TM1, TM5, TM6 and TM10 of the α -subunit, and thus clamps together the two halves of the α -subunit.

Comparison with the EM structure of the *E. coli* SecY complex

Functional interpretations of the X-ray structure must be based on the large body of experimental work carried out on systems from *E. coli*, mammals and *S. cerevisiae*; no functional data are available for archaeal SecY complexes, but the X-ray structure of the *M. jannaschii* complex is probably representative of all species. The amino-acid sequences of the *M. jannaschii* subunits are quite similar to those in eukaryotes and eubacteria ($\sim 50\%$ similarity) (Supplementary Fig. S1). Fitting the X-ray structure into the electron density map of the 2D crystal structure of the *E. coli* SecY complex, determined by EM¹⁹, shows that the TM segments of the SecY complexes from *M. jannaschii* and *E. coli* are arranged in nearly identical ways (Fig. 2a). We can identify all TMs in the *E. coli* α -subunit and observe that many features are conserved, including the position of the unusual segment following TM1 (Fig. 2a). The external loops between TM7 and TM8 and between TM5 and TM6 are shorter in eubacteria, causing small shifts in some helices, but the changes are confined largely to lipid-facing parts of the molecule, which have few conserved residues (Supplementary Fig. S6). The agreement between the X-ray and EM models also shows that the architecture is the same whether in detergent solution or in a phospholipid bilayer. Moreover, as the α -subunit in the 2D crystal is part of a dimer¹⁹ (Fig. 2a, b), oligomerization does not grossly change the conformation of the SecY complex.

The small subunits of the *E. coli* complex in the EM structure can

Table 1 Crystallographic statistics

Data set	SeMet (X25)*	Y1 mutant (8BM)*
Resolution (Å)	3.5	3.2
Unique reflections	14,439 (1,291)	18,118 (1,439)
I/ σ	30.9 (2.74)	26.9 (2.24)
Completeness (%)	98.4 (89.6)	93.2 (76.1)
$R_{\text{sym}}^\dagger$	0.08 (0.67)	0.05 (0.44)
Phasing to 3.8 Å		
FOM (from SOLVE)	0.29 (0.19)	
Map correlation ‡	0.24	
Mean phase difference ($^\circ$) §	75.1 (77.6)	
$R_{\text{crys}}^{\parallel}$	0.254 (0.383)	0.242 (0.414)
$R_{\text{free}}^{\parallel}$	0.334 (0.463)	0.287 (0.442)
r.m.s. deviation bond length (Å)	0.008	0.008
r.m.s. deviation bond angles ($^\circ$)	1.3	1.23
Mean B-factor	122.4	97.8

*Values in parentheses refer to data in the highest-resolution shell (3.63 Å to 3.50 Å and 3.31 Å to 3.20 Å in the wild-type and mutant data sets, respectively). SeMet, selenomethionine.

$^\dagger R_{\text{sym}} = \sum_{hkl} \sum_i |I(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I(hkl)$, where $\langle I(hkl) \rangle$ is the average intensity. All reflections were used.

‡ The map correlation coefficient is the correlation between a synthetic electron density map calculated on the basis of the final model and the map corresponding to the experimental set of phases, averaged over all grid points.

§ Mean phase difference between initial phases from SOLVE and phases calculated from the final refined wild-type model. The phase difference is defined as the average difference between experimental phases and phases calculated from the final refined model.

$^{\parallel} R_{\text{crys}} = \sum_{hkl} |F_{\text{obs}} - k|F_{\text{calc}}| / \sum_{hkl} F_{\text{obs}}$

$^{\parallel} R_{\text{free}}$ is the same as R_{crys} for a selected subset (10%) of the reflections that was not included in prior refinement calculations.

also be identified from a comparison with the *M. jannaschii* complex. The bacterial β -subunit (SecE) comprises two TMs, the second of which has the same position and orientation as the TM of Sec β of *M. jannaschii*²⁵ (Fig. 2a), suggesting that they may have analogous functions. The N-terminal TM of the bacterial β -subunit (SecE) has no correspondence in archaea and eukaryotes. The γ -subunit (SecE) of *E. coli* has two non-essential TMs at its N terminus²³, which correspond to the two helices that are far apart from the others in the 2D crystal structure of the *E. coli* protein (Fig. 2a). These helices are in approximately the same location as the N terminus of the *M. jannaschii* γ -subunit in the X-ray structure (Fig. 2b).

Translocation pore and plug

Experiments in different systems have shown that the SecY/SecE1p complex forms oligomers during translocation, but as we will argue later, our structure suggests that a single copy of the SecY/SecE1p complex serves as a functional channel. A large, funnel-like cavity with a diameter of 20–25 Å at the cytoplasmic side of the SecY complex could serve as a channel entrance (Fig. 3a, b). It contains many conserved residues (Supplementary Fig. S6), suggesting that it has an important function. The funnel tapers to a close in the middle of the membrane (Fig. 3b), indicating that the structure corresponds to a closed channel, impermeable to polypeptides or even small molecules.

How might the channel open and how could it recognize signals? A large body of data in the literature allows us to propose specific models for these and other properties. TM2a, which we call the 'plug', blocks the bottom of the funnel about halfway across the membrane, and separates the cytoplasmic side from the external aqueous space (Fig. 3a). We propose that the channel opens for

polypeptide translocation by displacement of the plug. Consistent with this hypothesis, a cysteine introduced into the plug at residue 67 of the *E. coli* α -subunit (SecY; corresponding to *M. jannaschii* Thr 61) can form a disulphide bridge *in vivo* with a cysteine introduced at residue 120 of the *E. coli* γ -subunit (SecE; *M. jannaschii* residue 64)²⁶. These residues are more than 20 Å apart in the closed state of the channel (Fig. 3a). Disulphide bridge formation results in a dominant-negative phenotype, as would be expected if the channel were locked into a permanently open state. Another combination of cysteines (*E. coli* residue 64 in the α -subunit (SecY) and 124 in the γ -subunit (SecE)) is also lethal, whereas combinations of neighbouring residues are not²⁶, suggesting that the helical structures of the plug and the γ -subunit are maintained during movement. Given that the TM of the γ -subunit is a continuous helix with one hydrophobic side, we assume that it remains stationary while the plug moves as a rigid body into a cavity on the external side of the channel next to the C terminus of the γ -subunit (Fig. 3b, c). This displacement requires a translation of ~22 Å towards the back of the molecule, as well as a shift of about 12 Å towards the external side of the membrane. The hinges for the motion could be the Gly residues at positions 49 and 68. Although not universally conserved, all species have small residues in this region that could serve this function. Movement of the plug would open the pore (Fig. 3b, c), resulting in an hourglass-shaped channel with aqueous funnels that taper to a constriction in the middle of the membrane (Fig. 3b).

The funnel-like cavities on both sides of the constriction would create an aqueous channel across the membrane (Fig. 4a, b). This feature is consistent with fluorescence life-time measurements³, which suggest that a translocating polypeptide is in an aqueous environment. The walls of both funnel-like cavities contain hydro-

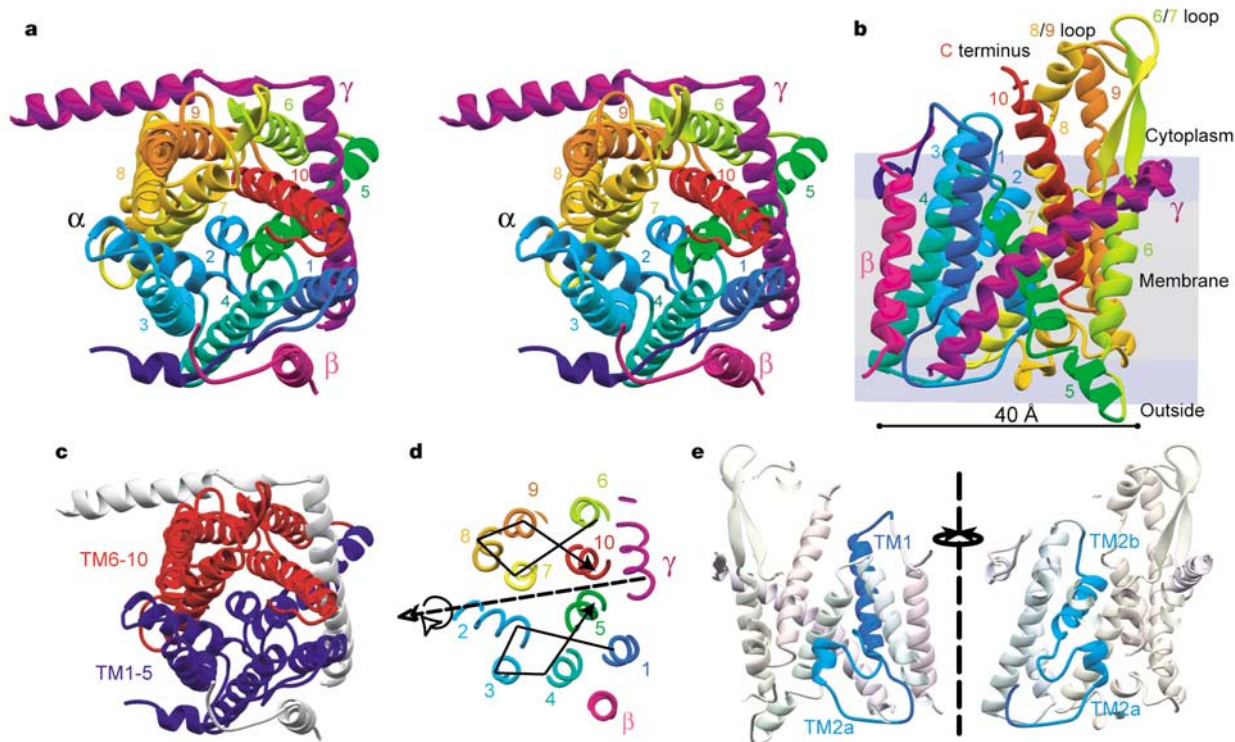


Figure 1 General architecture of the SecY complex. **a**, Stereo view from the cytosol (top). The α -subunit is coloured blue to red from the N to the C terminus with the TM segments numbered; the β -subunit is shown in pink and the γ -subunit in magenta. **b**, View from the back, with the phospholipid head group and hydrocarbon regions of the membrane shown in blue and grey in the background. Cytosolic loops probably involved in ribosome and SecA binding are indicated. **c**, Top view with the N- and C-terminal halves of the

α -subunit in blue and red, respectively. **d**, Top view sliced through the middle of the membrane. The solid lines connect the TMs from the N to the C terminus in the two halves; the dotted arrow shows the axis of internal symmetry. **e**, Slab views from the front and back, with the foreground removed and TM1 (dark blue), TM2a and TM2b (sky blue) highlighted.

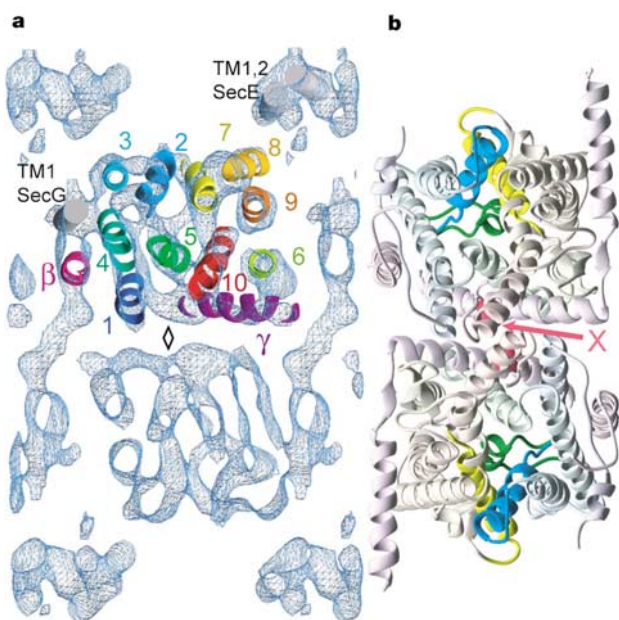


Figure 2 Comparison with the 2D crystal structure of the *E. coli* SecY complex. **a**, X-ray structure of the *M. jannaschii* SecY complex (coloured as in Fig. 1a), visually docked into the electron density map of the *E. coli* SecY complex, determined by cryo-electron microscopy of 2D crystals¹⁹. Shown is a 5 Å slab, viewed from the top, with the map contoured at 1 σ . The *M. jannaschii* TMs are numbered. TMs of the *E. coli* complex with no correspondence in *M. jannaschii* were fitted into the density as grey cylinders. The diamond symbol indicates the axis of two-fold symmetry in the *E. coli* complex. **b**, Modelled dimer of the SecY complex, in the same orientation as in **a**, with TM2b and TM7 coloured as in Fig. 1a. TM2a (plug) is shown in dark green. A cysteine introduced at the position indicated by a red sphere results in efficient crosslinks (X) between two γ -subunits.

philic residues (Fig. 4a, b), but there are few charges (not shown). The absence of charges is particularly remarkable for the cytoplasmic funnel; its interior is uncharged, but its outer rim in the cytoplasm contains a large number of both positive and negative residues.

The pore ring

The nature of the constriction suggests a mechanism by which the membrane barrier can be maintained during protein translocation. At its narrowest point, the channel is lined by a 'pore ring' of six

hydrophobic residues (Ile 75, Val 79, Ile 170, Ile 174, Ile 260 and Leu 406) (Fig. 3b, c). In *E. coli*, the corresponding residues are all isoleucines, and this amino acid is frequently found as pore residues in other species (Supplementary Fig. S1). The opening formed by the ring is about 5–8 Å. The hydrophobic residues may form a gasket-like seal around a translocating polypeptide, and bulky isoleucines may be particularly suitable seal residues. Different amino acids must pass through the pore ring, and the required flexibility may be provided largely by lateral shifts of the helices to which the pore residues are attached (see below). Although the seal is not likely to be perfect, it would significantly hinder the passage of small molecules during translocation. This model is different from previous proposals, in which the ribosome–channel junction and the binding of the luminal protein BiP provide the seal in co-translational translocation in mammals^{3,27}. Our model explains how the membrane barrier can be maintained in both co- and post-translational translocation pathways, and why a gap between the ribosome and the channel, seen in EM structures of the eukaryotic ribosome–channel complex^{15–18}, may not compromise the membrane barrier.

The hourglass shape of the open channel would minimize its interactions with a translocating polypeptide chain. Contacts would be confined largely to the constriction formed by the pore residues. Because the constriction consists of only a single layer of hydrophobic side chains (Fig. 3b), interactions with a polypeptide at its centre would probably be weak, even when hydrophobic amino acids are passing through. The lack of charges in the cavities on either side of the constriction may also help to minimize interactions between the channel and the translocating chain.

The diameter of the pore ring seen in the X-ray structure might be sufficient to accommodate an extended polypeptide chain, but with a little expansion it could allow passage of an α -helix. An increase in pore size could result from shifts in the helices that line the channel, perhaps facilitated by rearrangement of the loop between TM4 and TM5, in which the conserved Gly residues of a Gly-Ile-Gly-Ser-Gly-Ile-Gly sequence could serve as hinges. TM10 could also move, facilitated by conserved Gly residues in the membrane-embedded loop between TM9 and TM10, related by pseudo-symmetry to the 4/5 loop (Supplementary Fig. S5). These shifts could allow a variable pore width, and might explain how even bulky residues attached to the side chains of amino acids *in vitro*²⁸ or a disulphide-bonded polypeptide loop of 13 residues in a secretory protein²⁹ can be transferred through the channel. The pore size would effectively prevent the passage of folded domains, consistent with experimental data^{30,31}. Unless the pores of several copies of Sec61p/SecY complexes fuse into a larger pore (see below), the structure seems

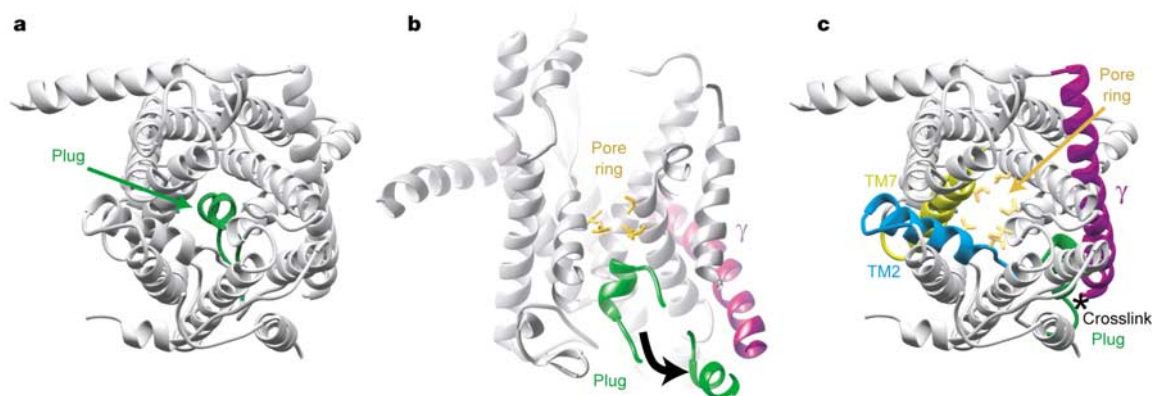


Figure 3 The channel pore. **a**, View from the top with TM2a (plug) coloured in dark green. **b**, View from the side with the front half of the model cut away. The modelled plug movement towards the γ -subunit (magenta) is indicated. The hydrophobic pore ring is

shown by the side chains coloured in gold. **c**, Top view with the plug modelled in its open position. TM2b and TM7 are coloured as in Fig. 1a. The star indicates the region where introduced cysteines result in crosslinking between TM2a and the TM of the γ -subunit.

to be inconsistent with the claim that the open channel has a diameter of at least 40 Å (ref. 32).

Signal-sequence binding and lateral gating

One trigger for channel opening is binding of the signal sequence of a translocation substrate^{3,33}. At the beginning of translocation of a secretory protein, the polypeptide inserts into the channel as a loop, with the signal sequence intercalated into the walls of the channel and the following, mature region located in the aqueous pore. Photocrosslinking experiments in a yeast *in vitro* system have shown that the hydrophobic region of a bound signal sequence forms a helix of approximately two turns, contacted on opposite sides by TM2 and TM7 (ref. 34). The part of the signal sequence preceding the hydrophobic core contacts TM8 (ref. 34). Each residue of the signal sequence can also be crosslinked to phospholipid molecules, indicating that the binding site is located at the interface between channel and lipid³⁴. The X-ray structure of the SecY complex enables us to rationalize these results. The signal-sequence intercalation site is located between TM2b and TM7 at the front of the cytoplasmic funnel (Figs 5a, b); it is not close to the TM of the γ -subunit, as previously proposed³⁴. The top of TM2b interacts with TM8, explaining why TM8 can be crosslinked to the N terminus of the bound signal sequence³⁴. In addition, the weak crosslinks of the signal sequence to TM2a (designated as TM1 in ref. 34) are also consistent with the structure. The residues forming the signal-sequence-binding site are well conserved (Supplementary Fig. S6). TM2b and TM7 are in opposite halves of the α -subunit, and intercalation of a signal sequence between them would require the front of the α -subunit to open. This could result from a small ($\sim 15^\circ$) hinge motion between TM5 and TM6 at the back of the molecule (Fig. 5), which would create a pore with dimensions of 15–20 Å front to back and 10–15 Å side to side, sufficient to allow loop insertion of a translocating polypeptide chain. As TM2b and

TM7 cross at an angle of about 70° , the signal sequence could not be in contact with both TMs for more than two turns, consistent with the minimal requirement for 6–7 consecutive hydrophobic residues in a functional signal sequence. Small variations in the hinge opening could permit a range of separations and relative orientations of TM2b and TM7, providing the flexibility needed to insert different signal sequences. Intercalation between TM2b and TM7 at the open front of the α -subunit would also explain why the signal sequence is exposed to lipid³⁴.

The location of the TM2b–TM7 intercalation site is consistent with a role for the signal sequence as a trigger for channel opening^{3,33}, because the signal sequence would have access to its binding site in the closed channel (Fig. 5b). We propose that the bound signal sequence destabilizes the interactions of the plug that keep it in the centre of the pore, opening the channel. Insertion into the pore of the polypeptide region distal to the signal sequence could fix the channel in the open state by preventing the plug from returning to its closed-state position. The plug could return and the pore would close only after the polypeptide chain had left the channel.

The TMs of nascent membrane proteins must exit laterally from the channel into the lipid phase. The X-ray structure shows that the front is the only place where the complex could open towards lipid; all other directions are blocked by the small subunits or obstructed by segments of the α -subunit. The lateral gate would be formed in the cytoplasmic half of the membrane by the interface between TM2b and TM8 and between TM2b and TM7, and in the external half by the interface between TM3 and TM7 (Fig. 5a, b). Opening the gate would require breaking of hydrophobic interactions and of hydrogen bonds between the conserved Gln 86 in TM2b and Thr 337 in TM8, and between Asn 268 in TM7, Glu 122 in TM3 and Thr 80 in TM2b (the last two are conserved; see Supplementary Fig. S1). Whereas signal sequences contain relatively short hydrophobic segments and may stay intercalated into the lateral gate until they are cleaved by signal peptidase, longer and more hydrophobic TMs may partition into lipid³⁵.

Signal-sequence suppressor mutations

Support for the postulated mechanism of pore opening comes from the location of signal-sequence suppressor (prl) mutations. Genetic studies in *E. coli* have shown that certain mutations in the SecY complex allow secretory proteins with defective or even deleted signal sequences to be transported^{36,37}. Most of the prl mutations are located in the centre of the channel, particularly on the internal side of TM7 and in the plug (Fig. 6). Inspection of specific mutations indicates that at least some of them could destabilize the closed state of the channel. For example, the mutations Phe64Cys and Asn65Tyr (prlA300 and prlA8914) in the *E. coli* α -subunit (SecY) correspond to changes in residues within the plug (Phe 58 and Trp 59) that face TM7. Other prl mutations change hydrophobic residues in the pore ring to hydrophilic residues (for example, Ile408Asn (*M. jannaschii* Leu 406) in prlA4). These mutations may stabilize the open state of the channel, in which the residues of the pore are in an aqueous environment, or they may facilitate widening of the pore during initiation of translocation. The prlG3 mutation Ser120Phe in the *E. coli* γ -subunit (SecE; *M. jannaschii* Gly 64) may also stabilize the open state of the channel, as this is the same residue that can be crosslinked to the plug. In general, our analysis supports the idea that prl mutations mimic the effect of signal-sequence binding—that is, they destabilize the closed state of the channel or stabilize the open state.

Binding of cytosolic channel partners

For translocation, the channel must associate with a partner, which can be—depending on the mode of translocation—the ribosome, SecA or the Sec62/63p complex. The ribosome must bind predominantly to the cytosolic domains in the C-terminal half of the α -subunit—that is, to the loops between TM6 and TM7 and between

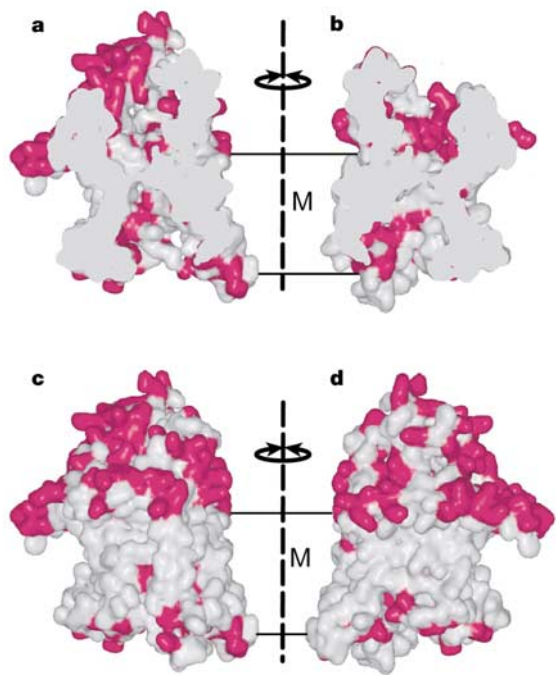


Figure 4 Distribution of polar residues. Accessible surface representations of the channel with the plug in its open-state position and polar amino-acid residues (S, T, C, H, D, E, N, Q, R, K) shown in red. **a, b**, Internal surface of the cytoplasmic and external funnels, revealed by cutting the complex into halves. The two views are related by 180° rotation. **c, d**, External surface of the complex shown from the front and back. The lines indicate the borders of the hydrophobic region of the membrane (M).

TM8 and TM9, and to the C-terminal tail (Fig. 1b). Proteolysis experiments with the mammalian Sec61p complex have shown that the loop between TM8 and TM9 and the C-terminal tail are required for high-affinity ribosome binding³⁸. The tip of the 8/9 loop contains several conserved residues, including the universally conserved Arg 360 and Arg 372, which are prime candidates for interaction with nucleic acid in the large ribosomal subunit³⁹. The cytosolic loops in the N-terminal half are much shorter, hardly protruding beyond the phospholipid-head-group region (Fig. 1b), and would not be expected to make strong contacts. Binding of the ribosome to only one half of the α -subunit would not prevent the relative movement of the two halves, which the structure suggests is required for signal-sequence intercalation. The protrusion of the cytosolic loops in the C-terminal half could place the binding platform for the ribosome 10–20 Å away from the plane of the membrane. A gap of this size has been seen by EM of ribosome-channel complexes^{15–18}.

In eubacteria, the same cytosolic domains that bind the ribosome probably interact with SecA. The loop between TM8 and TM9 is particularly important, and mutagenesis of residues corresponding to Arg 360 and a few residues following it results in translocation defects (for a review, see ref. 40). Mutations in the C-terminal tail, although not abolishing the interaction with SecA, prevent a conformational change in SecA that is required for translocation⁴⁰. Both the overlap of the interaction sites and size considerations suggest that the ribosome and SecA cannot bind at the same time. The narrow pore in the SecY complex does not explain observations that suggest a deep insertion of SecA into the channel^{11,41}.

Binding of a channel partner probably causes conformational changes in the Sec61p/SecY complex. This would be consistent with the observation that Sec61p channels with bound ribosomes have an increased conductance for small molecules compared with those without a channel partner or those containing a translocating polypeptide^{2,42}. Binding of a channel partner followed by intercalation of a signal sequence may be required to open the channel. In the case of a prl mutant in *E. coli*, SecA binding would be sufficient, and a signal sequence would not be needed.

Oligomerization of the Sec61/SecY complex

The translocating channel is likely to be an oligomer of the Sec61/SecY complex, containing between two and four copies^{15–19,43–45}, and it has been assumed that a large pore forms at the interface of several Sec61/SecY complexes. However, a side view shows that simple association of several copies of the complex could not form a hydrophilic pore in the membrane (Fig. 4c, d). A hydrophilic pore could be generated by multiple complexes only if they associated at their front surfaces and opened to fuse the α -subunits into a larger channel. The enlarged pore would be lined by the same residues as in a single complex, but would have a lateral exit site at each of the interfaces between α -subunits. Although we cannot exclude this possibility, we consider it unlikely. In fact, for the bacterial post-translational translocation system, experimental evidence suggests that a dimer with a back-to-back association of the complexes is the basic functional unit. This is the arrangement seen in the 2D crystal structure¹⁹ (Fig. 2). A cavity between the two monomers, previously thought to be a potential pore¹⁹, is entirely hydrophobic and is probably filled with lipids. A functional back-to-back association of two SecY complexes is supported by several different experiments. Cysteine crosslinking shows that two γ -subunits are close to one another during translocation⁴⁶, with the most efficient crosslinking occurring at residue 106 of the *E. coli* γ -subunit (SecE; *M. jannaschii* Ile 50), which is indeed at the interface between two γ -subunits in the 2D crystals (Fig. 2b). Crosslinking also shows that the N and C termini of two α -subunits are in proximity, and that the crosslinked dimer has translocation activity⁴⁷. A tandem molecule, in which the C terminus of the first α -subunit is linked with the N terminus of the second, is also functional⁴⁵. Because the lateral exit sides in the back-to-back dimer face different directions and there is no connection between the two pores (Fig. 2b), we conclude that one copy of the SecY complex is sufficient to serve as an active pore. In support of this proposal, a detergent-solubilized translocation intermediate contains just one copy of SecY complex associated with one SecA and one translocation substrate molecule^{45,48}. The active translocation channel may also be contained in a tetramer of SecY complexes⁴³, which could

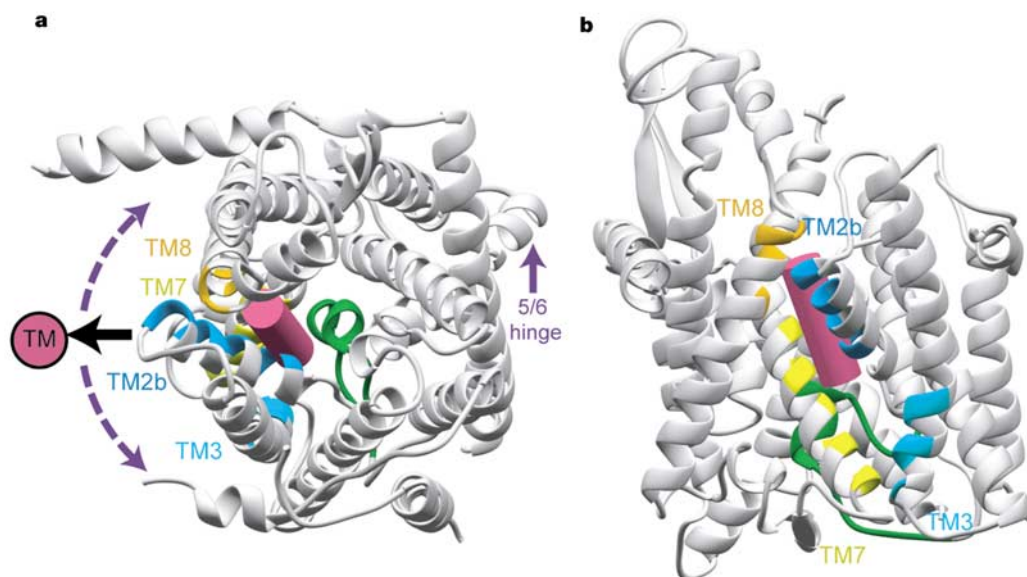


Figure 5 Signal-sequence-binding site and lateral gate. **a, b**, Views from the top (**a**) and the front (**b**), with faces of the helices that form the signal-sequence-binding site and the lateral gate through which TMs of nascent membrane proteins exit the channel into lipid highlighted in colours as in Fig. 1a. The plug is coloured in green. The hydrophobic core of the signal sequence probably forms a helix, modelled as a magenta cylinder, which

intercalates between TM2b and TM7 above the plug. Intercalation requires opening the front surface, as indicated by the broken arrows, with the hinge for the motion being the loop between TM5 and TM6 at the back of the molecule (5/6 hinge). A solid arrow pointing to the magenta circle in the top view indicates schematically how a TM of a nascent membrane protein would exit the channel into lipid.

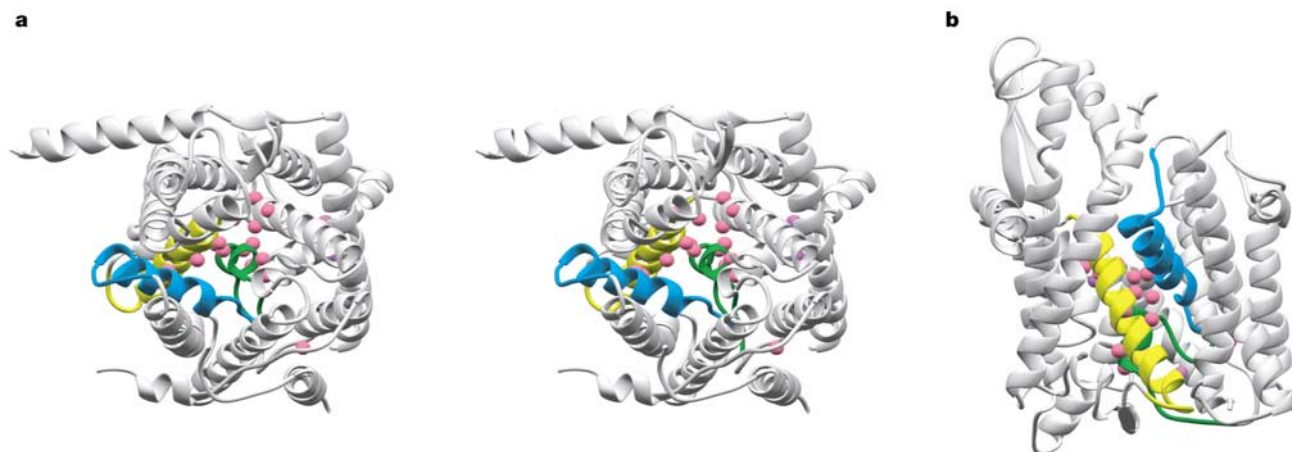


Figure 6 Signal-sequence suppressor (*prl*) mutations. **a**, Stereo view from the top with the C β atoms of the residues corresponding to *E. coli prl* mutations in the α -subunit (SecY)

and γ -subunit (SecE) shown as pink and purple spheres, respectively. TM2a, TM2b and TM7 are highlighted in colour. **b**, View from the front.

form by the association of two dimers side by side, similar to their arrangement in the 2D crystal lattice¹⁹. Crosslinks between the TM of the *E. coli* γ -subunit (SecE) and TM2 and TM7 of the α -subunit (SecY) have been observed⁴⁹, but they do not fit with either front-to-front or back-to-back orientation of the monomers.

Ribosome-associated Sec61p channels contain three or four copies of the complex^{15–18}. In view of their sequence similarity (Supplementary Fig. S1), it seems likely that the co-translational eukaryotic and the post-translational bacterial complexes function similarly. We therefore favour a model in which two dimers, with a back-to-back arrangement of the complexes, associate side by side beneath the ribosome. The front sides of all complexes would then face outwards, and only one complex at any given time would form the active pore and contain a translocating polypeptide. As the ribosome is asymmetric and would make different contacts with the four copies of the Sec61p/SecY complex, the monomers may have different conformations. Our model implies that the appearance of a pore in low-resolution EM structures of ribosome–channel

complexes is simply an indentation between the subunits rather than a channel. The recent higher-resolution EM structures do not have a pore, and the observed central indentation is offset from the exit site of the nascent chain from the ribosome^{17,18}.

Oligomers of the Sec61p complex may also be required for post-translational translocation in *S. cerevisiae*¹⁴, but there is no evidence that the signal sequence intercalates between TMs of different α -subunits⁵⁰. Thus, in all three translocation pathways, a single copy of the Sec61p/SecY complex in an oligomeric assembly may form the translocation pore.

If the active pore is formed by a monomer, what is the role of oligomerization? Allosteric interactions between the monomers, regulation of the binding of partners and display of sites for recruitment of other factors (for example, signal peptidase, oligosaccharide transferase, TRAM or TRAP) are possible answers. Nonetheless, the issue remains an important puzzle for future work.

A model for protein translocation

The X-ray structure allows us to propose the following refined model for the translocation of secretory proteins. Initially, the channel is closed because the plug blocks the pore (Fig. 7, stage 1). Next, a channel partner binds; depending on the mode of translocation and the organism, this can be a ribosome, the Sec62/63p complex or SecA (Fig. 7, stage 2 indicates the situation with a ribosome). Although part of an oligomer, only one copy of the SecY/Sec61p complex forms the active pore. The closed state of the channel may be destabilized by a conformational change, but binding of the partner alone is insufficient to completely open the channel. In the next step, the substrate inserts as a loop into the channel, with its signal sequence intercalated between TM2b and TM7, and with its mature region in the pore (Fig. 7, stage 3). Insertion requires a hinge motion to separate TM2b and TM7, and displacement of the plug to its open-state position close to the γ -subunit. The mature region of the polypeptide chain is then transported through the pore, and the signal sequence is cleaved at some point during translocation (Fig. 7, stage 4). While the polypeptide chain is moving from an aqueous cytoplasmic cavity to an external one, the pore ring forms a seal around the chain, hindering the permeation of other molecules. Finally, when the polypeptide has passed through, the plug returns to its closed-state position (Fig. 7, stage 5). Membrane protein biosynthesis, although less well understood, might occur in a similar way, except that TMs could move from the lateral gate all the way into lipid, and cytosolic domains would form by emerging through the gap between ribosome and channel. Several of these points differ from conventional

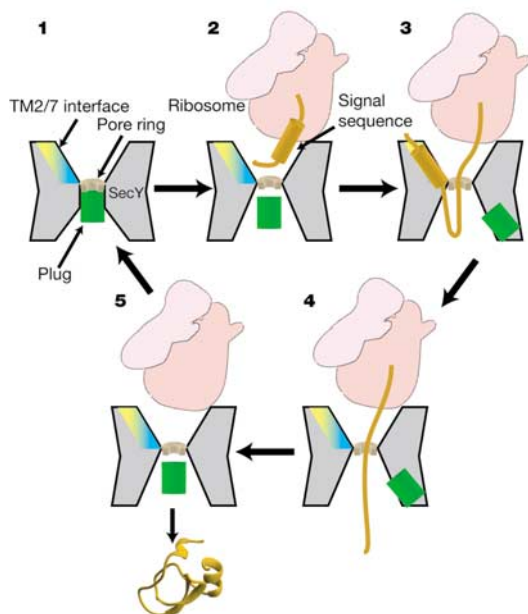


Figure 7 Different stages of translocation of a secretory protein. See main text for description.

models, but the structure now provides the basis for experimental testing. □

Methods

Protein preparation

We cloned SecY complexes from ten different species into the pBAD22 vector (see Supplementary Information) under control of the arabinose promoter. The plasmids contained the genes for the γ -, α - and β -subunits in this order, each with a separate translation-initiation site. The genes were amplified by polymerase chain reaction (PCR) from genomic DNA, and digested with *NcoI/SalI*, *SalI/XbaI* and *XbaI/KpnI*, respectively. Four-component ligation with *NcoI/KpnI*-digested pBAD22 resulted in the intergenic sequence AGGAGGAGCATC between the 5' restriction site and the initiation codon of the α - and β -subunits. The γ -subunit contained at its N terminus a hexa-histidine tag followed by a thrombin cleavage site, resulting after cleavage in the N-terminal sequence Gly-Ser instead of Met. The following procedure was used for the SecY complex from *M. jannaschii*, which gave the highest expression levels and was the most stable in a number of detergents. Initially, we crystallized the SecYE dimer, but the crystals never diffracted beyond ~ 6.5 Å. By consulting sequence databases, we discovered archaeal polypeptides with sequence similarity to the β -subunits in eukaryotes. We co-expressed the β -subunit from *M. jannaschii* and found that it associated with SecYE. The trimeric complex turned out to be significantly more stable than the dimer. For its purification, cells of *E. coli* strain C43 (DE3) were induced with 0.2% arabinose for 3 h at 37°C or 12–16 h at room temperature, and lysed on ice in TSG buffer (20 mM Tris, 300 mM NaCl, 10% glycerol, pH 7.8) using a microfluidizer. Membranes were collected by centrifugation for 40 min at 40,000 r.p.m., and solubilized in 1.5% (w/v) 1,2-diheptanoyl-*sn*-glycero-3-phosphocholine (DHPC; Avanti Polar Lipids) in TSG buffer. After centrifugation for 30 min at 40,000 r.p.m., the extract was loaded onto a 15–20 ml Ni column (metal-chelating Sepharose, Amersham Pharmacia Biotech). The column was washed with 20 volumes of 0.2% DHPC in TSG buffer in the presence of 30 mM imidazole. The SecY complex was eluted with 300 mM imidazole, exchanged into PBS (pH 7.4) containing 10% glycerol and 0.1% DHPC, and treated with bovine thrombin at 5–10 U mg⁻¹ protein for 16 h at 30°C. After adding imidazole to 30 mM, the solution was passed through a small (2 ml) Ni column. Further purification was carried out by gel filtration on Superdex-200 in buffer A (10 mM HEPES, pH 7.5, 100 mM NaCl, 10% glycerol, 0.1% DHPC), followed by ion-exchange chromatography on Mono-S (Pharmacia) using a gradient from 100 mM to 400 mM NaCl in buffer A. The β -subunit was present at sub-stoichiometric levels relative to the YE subunits; obtaining crystals depended critically on removing contaminating SecYE complexes from the trimeric complex, which was achieved by the gel-filtration and ion-exchange steps. The complex was concentrated to 10–15 mg ml⁻¹ with a Centrprep Plus-20 concentrator (Amicon; molecular mass cutoff, 30 kDa) and dialysed overnight at 4°C.

Selenomethionine-substituted SecY complex from *M. jannaschii* was expressed in wild-type C43 (DE3) cells by inhibition of the methionine biosynthesis pathway. The cells were grown in M9 minimal medium with 0.2% glucose and 5% (v/v) glycerol as carbon sources, and induced with 1% arabinose. Purification was as described for the native protein, with all buffers supplemented with 0.5 mM EDTA and 0.5 mM Tris-(2-carboxyethyl)phosphine to avoid oxidation.

Crystallization

Crystallization of the *M. jannaschii* trimeric SecY complex was performed by hanging-drop vapour diffusion at 4°C, mixing 1 μ l each of protein (5–10 mg ml⁻¹) and reservoir solutions. Initial crystallization conditions were found using a broad screen. After optimization, the best crystals were obtained with 50–55% PEG400, 50 mM glycine buffer, pH 9.0–9.5. They appeared overnight and grew to maximum dimensions of 150 $\times$ 150 $\times$ 400 μ m within a week. They belong to space group P2₁2₁2 and diffract to 3.4 Å. Selenomethionine-substituted crystals of the SecY complex were obtained with $\sim 45\%$ PEG400.

After obtaining an initial model, several double mutants were made in which residues involved in crystal contacts between the C terminus of one monomer and the 5/6 loop of another were changed. One of them (Lys422Arg/Val423Thr), called Y1, crystallized in thin plates at 35% PEG400.

Crystals were flash frozen in liquid nitrogen directly from the drop. We discovered that on gradually increasing the PEG concentration from 35–40% to 52%, the unit cells shrank substantially (Supplementary Table 1). A similar reduction in cell dimensions was observed on soaking certain heavy-atom derivatives such as K₂PtCl₆ into the crystals. The different crystal forms were useful for cross-crystal averaging of electron density maps.

Data collection, structure determination and refinement

Diffraction data were collected at 100 K on beamlines at either the National Synchrotron Light Source at Brookhaven National Labs (X25) or the Advanced Photon Source at Argonne National Labs (8BM, 14BM-C, 19ID) (see Table 1). The data were indexed and scaled with HKL2000. As the crystals were radiation sensitive, we collected highly redundant single-wavelength anomalous dispersion (SAD) data sets at the selenium peak wavelength. Thirteen out of 15 selenium sites were located using SOLVE. The selenium sites in other crystal forms were found by molecular replacement using AMORE. After solvent flattening with CNS, electron density maps were generated with useful phases to ~ 4 Å. An initial model was built with the program O for the α -helical TM segments, using selenium sites and large amino-acid side chains to determine the registry. The model was improved by an iterative process, using cross-crystal averaging, solvent flattening and histogram matching with a modified version of DMMULTI, and model building.

Molecular replacement with the improved model was used in AMORE to obtain electron

density maps for the native SecY complex and the Y1 mutant, which were then included in the averaging. Torsion angle refinement with hydrogen bond restraints and B-factor refinement of the model were performed with the Y1 data set using CNS, resulting in an R_{free} of 28.7%. This model was used in the subsequent refinement of the wild-type crystal form, giving an R_{free} of 33.4%. The following residues in the best-refined structure of the α -subunit were poorly defined, presumably because of conformational flexibility: 18–20, 48–55, 288–292 and 301–314. Figures shown in the paper were rendered with RIBBONS and POV-Ray (<http://www.povray.org/>), except for Fig. 4, which was generated with Raster3D and SPOCK, and Supplementary Fig. S4, which was made using Molscript.

Received 14 October; accepted 19 November 2003; doi:10.1038/nature02218.

Published online 3 December 2003.

- Matlack, K. E. S., Mothes, W. & Rapoport, T. A. Protein translocation—tunnel vision. *Cell* **92**, 381–390 (1998).
- Simon, S. M. & Blobel, G. A protein-conducting channel in the endoplasmic reticulum. *Cell* **65**, 371–380 (1991).
- Crowley, K. S., Liao, S. R., Worrell, V. E., Reinhart, G. D. & Johnson, A. E. Secretory proteins move through the endoplasmic reticulum membrane via an aqueous, gated pore. *Cell* **78**, 461–471 (1994).
- Rapoport, T. A., Jungnickel, B. & Kutay, U. Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.* **65**, 271–303 (1996).
- Mothes, W., Prehn, S. & Rapoport, T. A. Systematic probing of the environment of a translocating secretory protein during translocation through the ER membrane. *EMBO J.* **13**, 3937–3982 (1994).
- Brundage, L., Hendrick, J. P., Schiebel, E., Driessen, A. J. M. & Wickner, W. The purified *E. coli* integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. *Cell* **62**, 649–657 (1990).
- Gorlich, D. & Rapoport, T. A. Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell* **75**, 615–630 (1993).
- Deshaies, R. J., Sanders, S. L., Feldheim, D. A. & Schekman, R. Assembly of yeast Sec proteins involved in translocation into the endoplasmic reticulum into a membrane-bound multisubunit complex. *Nature* **349**, 806–808 (1991).
- Panzner, S., Dreier, L., Hartmann, E., Kostka, S. & Rapoport, T. A. Posttranslational protein transport in yeast reconstituted with a purified complex of Sec proteins and Kar2p. *Cell* **81**, 561–570 (1995).
- Matlack, K. E., Misselwitz, B., Plath, K. & Rapoport, T. A. BiP acts as a molecular ratchet during posttranslational transport of prepro- α -factor across the ER membrane. *Cell* **97**, 553–564 (1999).
- Economou, A. & Wickner, W. SecA promotes preprotein translocation by undergoing ATP-driven cycles of membrane insertion and deinsertion. *Cell* **78**, 835–843 (1994).
- Schiebel, E., Driessen, A. J. M., Hartl, F.-U. & Wickner, W. $\Delta\mu\text{H}^+$ and ATP function at different steps of the catalytic cycle of preprotein translocase. *Cell* **64**, 927–939 (1991).
- Irihimovitch, V. & Eichler, J. Post-translational secretion of fusion proteins in the halophilic archaea *Haloflex volcanii*. *J. Biol. Chem.* **278**, 12881–12887 (2003).
- Hanein, D. et al. Oligomeric rings of the Sec61p complex induced by ligands required for protein translocation. *Cell* **87**, 721–732 (1996).
- Beckmann, R. et al. Alignment of conduits for the nascent polypeptide chain in the ribosome-Sec61 complex. *Science* **19**, 2123–2126 (1997).
- Menetret, J. et al. The structure of ribosome-channel complexes engaged in protein translocation. *Mol. Cell* **6**, 1219–1232 (2000).
- Beckmann, R. et al. Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372 (2001).
- Morgan, D. G., Menetret, J. F., Neuhof, A., Rapoport, T. A. & Akey, C. W. Structure of the mammalian ribosome-channel complex at 17 Å resolution. *J. Mol. Biol.* **324**, 871–886 (2002).
- Breyton, C., Haase, W., Rapoport, T. A., Kuhlbrandt, W. & Collinson, I. Three-dimensional structure of the bacterial protein-translocation complex SecYEG. *Nature* **418**, 662–665 (2002).
- Murata, K. et al. Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599–605 (2000).
- Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. X-ray structure of a CIC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* **415**, 287–294 (2002).
- Flower, A. M., Osborne, R. S. & Silhavy, T. J. The allele-specific synthetic lethality of prlA-prlG double mutants predicts interactive domains of SecY and SecE. *EMBO J.* **14**, 884–893 (1995).
- Murphy, C. K. & Beckwith, J. Residues essential for the function of SecE, a membrane component of the *Escherichia coli* secretion apparatus, are located in a conserved cytoplasmic region. *Proc. Natl. Acad. Sci. USA* **91**, 2557–2561 (1994).
- Satoh, Y., Mori, H. & Ito, K. Nearest neighbor analysis of the SecYEG complex. 2. Identification of a SecY-SecE cytosolic interface. *Biochemistry* **42**, 7442–7447 (2003).
- Nishiyama, K., Suzuki, T. & Tokuda, H. Inversion of the membrane topology of SecG coupled with SecA-dependent preprotein translocation. *Cell* **85**, 71–81 (1996).
- Harris, C. R. & Silhavy, T. J. Mapping an interface of SecY (PrlA) and SecE (PrlG) by using synthetic phenotypes and *in vivo* cross-linking. *J. Bacteriol.* **181**, 3438–3444 (1999).
- Hamman, B. D., Hendershot, L. M. & Johnson, A. E. BiP maintains the permeability barrier of the ER membrane by sealing the luminal end of the translocon pore before and early in translocation. *Cell* **92**, 747–758 (1998).
- Kurzchalia, T. V. et al. tRNA-mediated labelling of proteins with biotin. A nonradioactive method for the detection of cell-free translation products. *Eur. J. Biochem.* **172**, 663–668 (1988).
- Tani, K., Tokuda, H. & Mizushima, S. Translocation of proOmpA possessing an intramolecular disulfide bridge into membrane vesicles of *Escherichia coli*. Effect of membrane energization. *J. Biol. Chem.* **265**, 17341–17347 (1990).
- Mingarro, L., Nilsson, I., Whitley, P. & von Heijne, G. Different conformations of nascent polypeptides during translocation across the ER membrane. *BMC Cell Biol.* **1**, 3 (2000).
- Kowarik, M., Kung, S., Martoglio, B. & Helenius, A. Protein folding during cotranslational translocation in the endoplasmic reticulum. *Mol. Cell* **10**, 769–778 (2002).
- Hamman, B. D., Chen, J. C., Johnson, E. E. & Johnson, A. E. The aqueous pore through the translocon has a diameter of 40–60 Å during cotranslational protein translocation at the ER membrane. *Cell* **89**, 535–544 (1997).
- Jungnickel, B. & Rapoport, T. A. A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. *Cell* **82**, 261–270 (1995).

34. Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94**, 795–807 (1998).
35. Heinrich, S. U., Mothes, W., Brunner, J. & Rapoport, T. A. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **102**, 233–244 (2000).
36. Bieker, K. L., Phillips, G. J. & Silhavy, T. J. The sec and prl genes of *Escherichia coli*. *J. Bioenerg. Biomembr.* **22**, 291–310 (1990).
37. Derman, A. L., Puziss, J. W., Bassford, P. J. & Beckwith, J. A signal sequence is not required for protein export in prlA mutants of *Escherichia coli*. *EMBO J.* **12**, 879–888 (1993).
38. Raden, D., Song, W. & Gilmore, R. Role of the cytoplasmic segments of Sec61 α in the ribosome-binding and translocation-promoting activities of the Sec61 complex. *J. Cell Biol.* **150**, 53–64 (2000).
39. Prinz, A., Behrens, C., Rapoport, T. A., Hartmann, E. & Kalies, K. U. Evolutionarily conserved binding of ribosomes to the translocation channel via the large ribosomal RNA. *EMBO J.* **19**, 1900–1906 (2000).
40. Mori, H. & Ito, K. The Sec protein-translocation pathway. *Trends Microbiol.* **9**, 494–500 (2001).
41. Kim, Y. J., Rajapandi, T. & Oliver, D. SecA protein is exposed to the periplasmic surface of the *E. coli* inner membrane in its active state. *Cell* **78**, 845–853 (1994).
42. Heritage, D. & Wonderlin, W. F. Translocon pores in the endoplasmic reticulum are permeable to a neutral, polar molecule. *J. Biol. Chem.* **276**, 22655–22662 (2001).
43. Manting, E. H., van Der Does, C., Remigy, H., Engel, A. & Driessen, A. J. SecYEG assembles into a tetramer to form the active protein translocation channel. *EMBO J.* **19**, 852–861 (2000).
44. Mori, H. *et al.* Fluorescence resonance energy transfer analysis of protein translocase. SecYE from *Thermus thermophilus* HB8 forms a constitutive oligomer in membranes. *J. Biol. Chem.* **278**, 14257–14264 (2003).
45. Duong, F. Binding, activation and dissociation of the dimeric SecA ATPase at the dimeric SecYEG translocase. *EMBO J.* **22**, 4375–4384 (2003).
46. Kaufmann, A., Manting, E. H., Veenendaal, A. K., Driessen, A. J. & van der Does, C. Cysteine-directed cross-linking demonstrates that helix 3 of SecE is close to helix 2 of SecY and helix 3 of a neighboring SecE. *Biochemistry* **38**, 9115–9125 (1999).
47. van der Sluis, E. O., Nouwen, N. & Driessen, A. J. SecY–SecY and SecY–SecG contacts revealed by site-specific crosslinking. *FEBS Lett.* **527**, 159–165 (2002).
48. Yahr, T. L. & Wickner, W. T. Evaluating the oligomeric state of SecYEG in preprotein translocase. *EMBO J.* **19**, 4393–4401 (2000).
49. Veenendaal, A. K., van der Does, C. & Driessen, A. J. Mapping the sites of interaction between SecY and SecE by cysteine scanning mutagenesis. *J. Biol. Chem.* **276**, 32559–32566 (2001).
50. Plath, K., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Interactions between Sec-complex and prepro- α -factor during posttranslational protein transport into the ER. *Mol. Biol. Cell* (in the press).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. MacKinnon for advice and suggestions of reagents; F. Duong for clones; J. Walker for C43 cells; C. Vornrhein, T. Terwilliger and K. Cowtan for help with software; M. Becker, L. Berman and S. LaMarra for support at beamline X25 at the National Synchrotron Light Source (Brookhaven National Laboratory, supported by the US Department of Energy, Division of Materials Sciences and Division of Chemical Sciences); A. Joachimiak, S. Ginell and R. Alkire for help at beamline 19ID at the Advanced Photon Source (supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences); and C. Ogata and M. Capel for help at beamline 8BM at the Advanced Photon Source (Northeastern Collaborative Access Team supported by an award from the National Center for Research Resources at the National Institutes of Health). We thank C. Akey, V. Ramakrishnan and particularly K. Matlack for critical reading of the manuscript. This work was supported by a fellowship from the Damon Runyon Cancer Research Foundation to W.M.C., and by fellowships from the Human Frontier Science Program Organization to I.C. and Y.M. E.H. was supported by grants from the Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie. T.A.R. and S.C.H. are Howard Hughes Medical Institute Investigators.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.A.R. (tom_rapoport@hms.harvard.edu). Coordinates for the native and mutant SecY complexes have been deposited in the Protein Data Bank, accession codes 1RHZ and 1RH5, respectively.

A correlation between the cosmic microwave background and large-scale structure in the Universe

Stephen Boughn¹ & Robert Crittenden²

¹Department of Astronomy, Haverford College, Haverford, Pennsylvania 19041, USA

²Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 2EG, UK

Observations of distant supernovae and the fluctuations in the cosmic microwave background (CMB) indicate that the expansion of the Universe may be accelerating¹ under the action of a 'cosmological constant' or some other form of 'dark energy'. This dark energy now appears to dominate the Universe and not only alters its expansion rate, but also affects the evolution of fluctuations in the density of matter, slowing down the gravitational collapse of material (into, for example, clusters of galaxies) in recent times. Additional fluctuations in the temperature of CMB photons are induced as they pass through large-scale structures² and these fluctuations are necessarily correlated with the distribution of relatively nearby matter³. Here we report the detection of correlations between recent CMB data⁴ and two probes of large-scale structure: the X-ray background⁵ and the distribution of radio galaxies⁶. These correlations are consistent with those predicted by dark energy, indicating that we are seeing the imprint of dark energy on the growth of structure in the Universe.

In the standard model of the origin of structure, most of the fluctuations in the CMB were imprinted as the photons last scattered off free electrons, when the universe was just 400,000 years old (at a redshift of $z \approx 1,100$). However, once the dark energy (or the spatial curvature⁷) becomes important dynamically ($z \approx 1$),

additional fluctuations are induced in the CMB photons by what is known as the integrated Sachs–Wolfe (ISW) effect². The gravitational potentials of large, diffuse concentrations and rarefactions of matter begin to evolve, causing the energy of photons passing through them to change by an amount that depends on the depth of the potentials. The amplitude of these ISW fluctuations tends to be small compared to the fluctuations originating at the epoch of last scattering except on very large scales. However, as ISW fluctuations were created more locally, it is expected that the CMB fluctuations should be partially correlated with galaxies, which serve as tracers of the large-scale matter distribution.

Detecting the relatively weak ISW effect requires correlating the CMB with the distribution of galaxies spread over a large fraction of the observable Universe. This necessitates a survey of galaxies covering much of the sky, out to distances of many billions of light years ($z \approx 1$). Focus thus has been on luminous active galaxies, which are believed to trace the mass distribution on large scales. Although active galaxies emit at a wide range of frequencies, the most useful maps are in the hard X-rays (2–10 keV), where they dominate the X-ray sky, and in the radio, where the number counts are dominated by sources at $z \lesssim 1$. The full sky map of the intensity of the hard X-ray background made by the HEAO-1 satellite⁵ and a map of the number density of radio sources provided by the NRAO VLA sky survey (NVSS)⁶ are two of the best maps in which to look for this effect. To predict the expected level of the ISW effect in these two surveys, it is essential to know both the inherent clustering of the sources and their distribution in redshift. The former can be determined from previous studies of the X-ray and radio auto-correlation functions^{8,9} and the latter can be estimated from deep, pointed surveys^{10,11}.

Previous attempts at correlating the CMB with both the HEAO-1 and the NVSS maps have yielded only upper limits^{8,9,12}. Here we repeat these analyses using the much improved CMB maps recently provided by the WMAP satellite mission⁴. We have compared two different CMB maps generated from this data (the 'internal linear

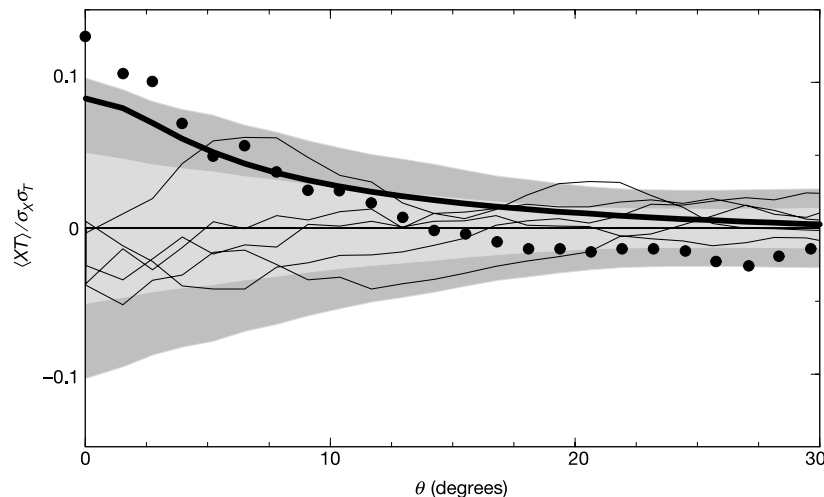


Figure 1 The X-ray background fluctuations are correlated with the microwave sky at a higher level than would be expected by chance. Here we plot the cross correlation (filled circles) between the X-ray intensity (X) measured by HEAO-A1 and the CMB temperature fluctuations (T) measured by WMAP. We normalize by the r.m.s. levels of the two maps (σ_X and σ_T , respectively), which makes it independent of any linear biasing of the survey. To give an idea of the level of accidental correlations, the shaded areas show the 1σ and 2σ regions derived from simulated, uncorrelated CMB maps with the same power spectrum as the WMAP data. The thin solid lines show the results for five such Monte Carlo simulations, and we see that the signals in neighbouring bins are highly correlated for a given realization. The signal-to-noise ratio is greatest at smaller angular separations,

even though the error is amplified because fewer pairs of pixels contribute to the correlation. For $\theta = 0^\circ, 1.3^\circ$ and 2.6° , the Monte Carlo trials exceed the amplitude of the actual X-ray/CMB correlation only 0.3%, 0.8% and 0.3% of the time, respectively. The bold line shows theoretical predictions for the ISW effect in a cosmological-constant-dominated Universe, using the best-fit WMAP model for scale-invariant fluctuations¹⁹ (dark energy fraction $\Omega_\Lambda = 0.73$, Hubble constant $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$ with $h = 0.72$, matter density $\Omega_m h^2 = 0.14$ and baryon density $\Omega_b h^2 = 0.024$.) At larger angular separations, the observed correlations appear to fall faster than predicted by theory, but the low signal-to-noise ratio makes it difficult to say whether this is a real effect.

combination' map¹³, and the 'cleaned' map of ref. 14) with the HEAO and NVSS maps. The dominant 'noise' in the maps is due to the fluctuations in the CMB itself and is well characterized, while the instrument noise is negligible on the angular scales relevant to the present analysis. To reduce possible contamination by emission from our Milky Way Galaxy as well as from other nearby radio sources, these maps were masked with the most aggressive foreground template provided by the WMAP team¹³. The X-ray map was similarly masked and was corrected for several large-scale systematics including a linear drift of the detectors⁹. Finally, the NVSS catalogue was corrected for a systematic variation of source counts with declination⁸. The sky coverage was 68% for the CMB maps, 56% for the NVSS map, and 33% for the X-ray map. The reduced coverage of the X-ray map was the result of its low resolution (3°) and a larger number of foreground sources. Significantly, however, our results do not depend on the corrections or the level of masking of the maps.

A standard measure of the correspondence of two data sets is the cross-correlation function, $CCF(\theta)$. It represents the extent to which the two measures of the sky separated by an angle θ are correlated. Figures 1 and 2 show the CCFs of the WMAP data with the X-ray and radio surveys. The results using the 'internal linear combination' and 'cleaned' CMB maps were entirely consistent with each other, and we plot the averages of the results for the two maps. Four hundred Monte Carlo simulations of the CMB sky were also cross-correlated with the actual X-ray and radio maps. The r.m.s. values of these trials are taken as the errors for each bin and these are highly correlated owing to the large-scale structure in the maps. For comparison, the uncertainties were also estimated from the data themselves by rotating the maps with respect to each other, and the errors determined in this way were very similar to those found using the simulated maps. The significance of the detection of the ISW effect is estimated by fitting the amplitude of the theoretical profile to the data. For the X-ray/CMB CCF, the correlations are detected at a confidence level (CL) of 99.9%, that is, a 3.0σ detection. The radio/CMB CCF is detected at a 99.4% CL, or a 2.5σ detection. These confidence limits are consistent with the number of times the $CCF(\theta)$ of the Monte Carlo trials exceed the values of the measured $CCF(\theta)$.

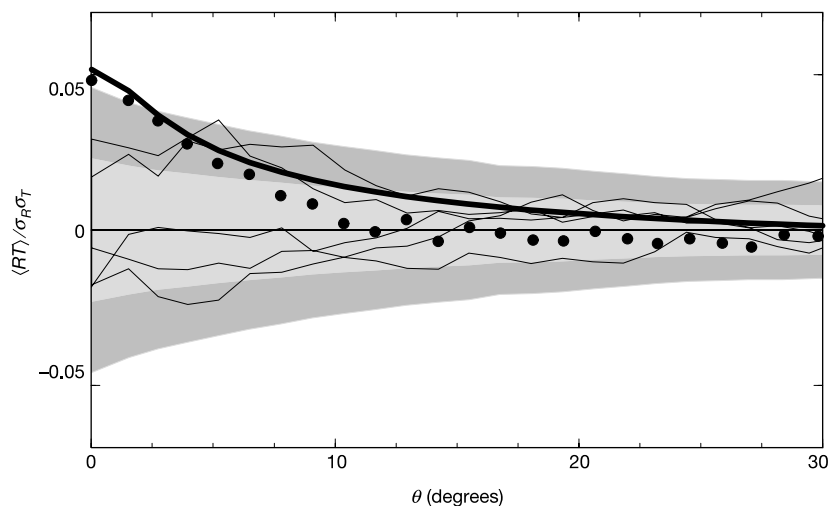


Figure 2 The distribution of radio galaxies is correlated with the microwave sky. Here we plot the correlation (filled circles) between the NVSS radio galaxy number counts (R) and the WMAP temperature maps (T). The other curves are as in Fig. 1. The Monte Carlo trials exceed the amplitude of the actual radio/CMB correlation in the lowest three bins 1.2%, 1.9% and 3.4% of the time, respectively. Again, there is good agreement with the theoretical predictions, with the signal again falling off somewhat faster than predicted at larger angles. The lower amplitude compared to the X-ray result is due to the higher resolution of the radio map, which increases the radio r.m.s. (σ_R) without significantly

The theoretical curves for the ISW effect also are shown in Figs 1 and 2, and fall off with angular separation similarly to the observed correlations. For these calculations, we have assumed that the fluctuations in matter density, $\delta\rho$, are traced directly by the fluctuations in the number density of the sources. That is, $\delta N/N = b\delta\rho/\rho$ where the bias factor, b , is assumed to be constant and its value is derived from the auto-correlation function of the X-ray or radio sources. The predictions assume a scale-invariant primordial spectrum and the best-fit cosmology determined from WMAP. Although the derived bias factors depend on precisely how the sources are assumed to be distributed in redshift, the expected cross-correlation is only weakly dependent on this or on the possible evolution of the bias factor with redshift. Detailed considerations for calculating the amplitude of the ISW effect can be found elsewhere^{3,12}.

It is important to consider what else might be contributing to the observed correlation. Unmasked foreground sources are not a major contaminant, because the observed correlations do not change when the Galactic cuts are varied nor when the aggressiveness of the masking is changed. In fact, the amplitudes of the CCFs are essentially unchanged when no masking at all is applied to the CMB maps. Also, because the results from analysing the different hemispheres of the maps separately are consistent, it is unlikely that a small number of diffuse foreground sources is responsible for the signal. Fluctuations caused by the Sunyaev-Zeldovich effect—the scattering of CMB photons as they travel through the ionized intergalactic medium of rich clusters of galaxies—are expected to contribute at smaller angles and be anti-correlated with the matter distribution¹⁵. Finally, microwave emission from the radio/X-ray sources could also produce correlations, but on much smaller angular scales than the observed signal.

The correlated signal evident in the figures is large enough to have been seen in our previous analyses using a COBE satellite combination map^{8,9}, albeit with a smaller signal-to-noise ratio because of that map's significant instrument noise and low angular resolution. However, we found no evidence for a correlation in the COBE analysis. When the differences in angular resolution are taken into account, the WMAP and COBE maps are consistent given COBE's

increasing the cross-correlation. The consistency of the NVSS and X-ray CCFs suggests that the signal is not the result of unknown systematics in either the X-ray or the NVSS map. It is possible that microwave emission from the radio/X-ray sources themselves could result in the positive correlation of the maps. However, extrapolations of the frequency spectra of the radio galaxies indicate that the microwave emission is much smaller than the observed signal⁶. In addition, the clustering of radio sources is on a much smaller angular scale than the apparent signals, so any such contamination would appear only at $\theta \approx 0^\circ$ and not at larger angles.

large instrumental noise ($\sim 70 \mu\text{K}$ per pixel)⁴. The difference between the observed cross-correlations implies that a small part (r.m.s. $\approx 1\text{--}2 \mu\text{K}$) of the (COBE-WMAP) difference map is anti-correlated with the X-ray and radio maps. This is an unlikely occurrence (1 in 100) if the COBE instrument noise model is correct. There are known systematics in the COBE data¹⁶ above this level, but the large instrument noise makes it difficult to ascertain whether this is the origin of the problem. In any case, the source of the discrepancy almost certainly lies with the COBE data, which contain much larger noise and systematic effects than the WMAP data. Thus, we believe the WMAP correlations presented here are robust. Further discussion of this and other issues are available as Supplementary Information.

The confidence levels that we quote (99.9% and 99.4%) are not primarily limited by experimental error, but by the fact that we can only observe a finite region of the Universe. We find it significant that both signals are consistent with the theoretical predictions with no free parameters. However, it should be emphasized that these two measurements are by no means independent. Both of them use the same CMB maps and, furthermore, the X-ray background is highly correlated with the NVSS radio sources¹⁷. If the distribution of sources of one of the maps by chance coincided with the fluctuations of the CMB, one would expect the other map also to be correlated with the CMB to some degree. The coincidence of the expected amplitudes of the two signals is encouraging but by no means definitive. On the other hand, the detection of the ISW-type signal in both CCFs gives a strong indication that they are not due to unknown systematic effects in the maps. The radio and X-ray data were gathered by quite distinct methods, and it would be surprising if unknown systematics in the two maps were correlated in any way. We will present more detailed analysis of these issues elsewhere.

We conclude that we have observed the ISW effect. If so, these observations offer the first direct glimpse into the production of CMB fluctuations, and provide important, independent confirmation of the new standard cosmological model: an accelerating universe, dominated by dark energy. It should be pointed out that measurements of the power spectrum of CMB fluctuations do not show evidence of increased power on large angular scales ($\theta > 20^\circ$) as predicted by the ISW effect, but rather indicate that there may be power missing⁴. Although this deficit is only at the 2σ level, it is intriguing, and may be telling us something about the formation of the very largest structures in the Universe. The consequences of the ISW effect reported here are primarily on intermediate angular scales, and are not in direct conflict with the power deficit on larger angular scales. Finally, we note that the WMAP-NVSS results have recently been independently analysed by the WMAP team, who find a similar level of correlation¹⁸. □

Received 15 April; accepted 8 October 2003; doi:10.1038/nature02139.

1. Bahcall, N., Ostriker, J. P., Perlmutter, S. & Steinhardt, P. The cosmic triangle: Revealing the state of the universe. *Science* **284**, 1481–1488 (1999).
2. Sachs, R. K. & Wolfe, A. M. Perturbations of a cosmological model and angular variations of the microwave background. *Astrophys. J.* **147**, 73–90 (1967).
3. Crittenden, R. & Turok, N. Looking for a cosmological constant with the Rees-Sciama effect. *Phys. Rev. Lett.* **76**, 575–578 (1996).
4. Bennett, C. L. *et al.* First year Wilkinson Microwave Anisotropy Probe observations: Preliminary maps and basic results. *Astrophys. J. Suppl.* **148**, 1–27 (2003).
5. Boldt, E. The cosmic X-ray background. *Phys. Rep.* **146**, 215–257 (1987).
6. Condon, J. *et al.* The NRAO VLA sky survey. *Astron. J.* **115**, 1693–1716 (1998).
7. Kinkhabwala, A. & Kamionkowski, M. New constraint on open cold-dark-matter models. *Phys. Rev. Lett.* **82**, 4172–4175 (1999).
8. Boughn, S. & Crittenden, R. Cross correlation of the CMB with radio sources: Constraints on an accelerating universe. *Phys. Rev. Lett.* **88**, 021302 (2002).
9. Boughn, S., Crittenden, R. & Koehn, G. The large scale structure of the X-ray background and its cosmological implications. *Astrophys. J.* **580**, 672–684 (2002).
10. Cowie, L. L., Barger, A. J., Bautz, M. W., Brandt, W. N. & Garmire, G. P. The redshift evolution of the 2–8 keV X-ray luminosity function. *Astrophys. J.* **584**, 57–60 (2003).
11. Dunlop, J. S. & Peacock, J. A. The redshift cut-off in the luminosity function of radio galaxies and quasars. *Mon. Not. R. Astron. Soc.* **247**, 19–42 (1990).
12. Boughn, S., Crittenden, R. & Turok, N. Correlations between the cosmic X-ray and microwave

backgrounds: Constraints on a cosmological constant. *New Astron.* **3**, 275–291 (1998).

13. Bennett, C. L. *et al.* First year Wilkinson Microwave Anisotropy Probe observations: Foreground emission. *Astrophys. J. Suppl.* **148**, 97–117 (2003).
14. Tegmark, M., de Oliveira-Costa, A. & Hamilton, A. A high resolution foreground cleaned CMB map from WMAP. Preprint at (<http://xxx.lanl.gov/astro-ph/0302496>) (2003).
15. Peiris, H. V. & Spergel, D. N. Cross-correlating the Sloan Digital Sky Survey with the microwave sky. *Astrophys. J.* **540**, 605–613 (2000).
16. Kogut, A. *et al.* Calibration and systematic error analysis for the COBE DMR 4 year sky maps. *Astrophys. J.* **470**, 653–673 (1996).
17. Boughn, S. Cross-correlation of the 2–10 keV X-ray background with radio sources: Constraining the large-scale structure of the X-ray background. *Astrophys. J.* **499**, 533–541 (1998).
18. Nolte, M. R. *et al.* First year Wilkinson Microwave Anisotropy Probe observations: Dark energy induced correlation with radio sources. Preprint at (<http://xxx.lanl.gov/astro-ph/0305907>) (2003).
19. Spergel, D. N. *et al.* First year Wilkinson Microwave Anisotropy Probe observations: Determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to M. Nolte, L. Page and the rest of the WMAP team, as well as to N. Turok, B. Partridge and B. Bassett, for useful conversations. R.C. acknowledges financial support from a PPARC fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.C. (Robert.Crittenden@port.ac.uk).

A large population of ‘Lyman-break’ galaxies in a protocluster at redshift $z \approx 4.1$

George K. Miley¹, Roderik A. Overzier¹, Zlatan I. Tsvetanov², Rychard J. Bouwens³, Narciso Benítez², John P. Blakeslee², Holland C. Ford², Garth D. Illingworth³, Marc Postman⁴, Piero Rosati⁵, Mark Clampin⁴, George F. Hartig⁴, Andrew W. Zirm¹, Huub J. A. Röttgering¹, Bram P. Venemans¹, David R. Ardila², Frank Bartko⁶, Tom J. Broadhurst⁷, Robert A. Brown², Chris J. Burrows², E. S. Cheng⁸, Nicholas J. G. Cross², Carlos De Breuck⁵, Paul D. Feldman², Marijn Franx¹, David A. Golimowski², Caryl Gronwall², Leopoldo Infante⁹, André R. Martel², Felipe Menanteau², Gerhardt R. Meurer², Marco Sirianni², Randy A. Kimble⁸, John E. Krist⁶, William B. Sparks⁴, Hien D. Tran², Richard L. White⁴ & Wei Zheng²

¹Leiden Observatory, University of Leiden, PO Box 9513, Leiden, 2300 RA, The Netherlands

²Department of Physics & Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA

³Lick Observatory, University of California, Santa Cruz, California 95064, USA

⁴Space Telescope Science Institute, Baltimore, Maryland 21218, USA

⁵European Southern Observatory, Garching, D-85748, Germany

⁶Bartko Science & Technology, Mead, Colorado 80542-0670, USA

⁷The Racah Institute of Physics, Hebrew University, Jerusalem, 91904, Israel

⁸NASA-Goddard Space Flight Centre, Greenbelt, Maryland 20771, USA

⁹Departamento de Astronomía y Astrofísica, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile

The most massive galaxies and the richest clusters are believed to have emerged from regions with the largest enhancements of mass density^{1–4} relative to the surrounding space. Distant radio galaxies may pinpoint the locations of the ancestors of rich clusters, because they are massive systems associated with ‘overdensities’ of galaxies that are bright in the Lyman- α line of hydrogen^{5–7}. A powerful technique for detecting high-redshift galaxies is to search for the characteristic ‘Lyman break’ feature in the galaxy colour, at wavelengths just shortwards of Ly α , which is due to absorption of radiation from the galaxy by the

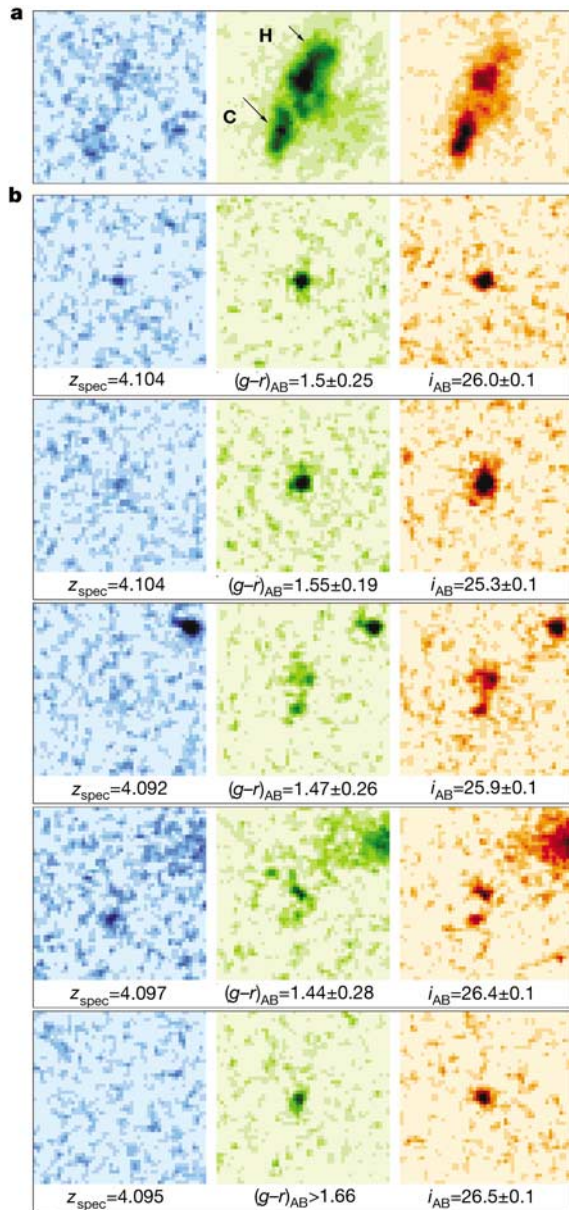


Figure 1 Deep images of Ly α -emitting protocluster galaxies. Images show galaxy morphologies observed through three filters: g band (left), r band (middle) and i band (right). Each $2.5'' \times 2.5''$ image has been smoothed by a gaussian function with a full-width at half-maximum of 1.5 pixels ($0.074''$). The observations were carried out between 8 and 12 July 2002 with the Wide Field Channel of the ACS²¹. The total observing time of 13 orbits was split over the broad-band filters F475W (g band, four orbits), F625W (r band, four orbits) and F775W (i band, five orbits), thereby bracketing redshifted Ly α at $6,214 \text{ \AA}$. During each orbit, two 1,200-s exposures were made, to facilitate the removal of cosmic rays. The observations were processed through the ACS GTO pipeline²⁶ to produce registered, cosmic-ray-rejected images. The limiting 2σ magnitudes in a 0.2-arcsec^2 aperture were 28.71 (F475W), 28.44 (F625W) and 28.26 (F775W). Object detection and photometry were then obtained using SExtractor²⁷. **a**, The clumpy radio galaxy TN J1338-1942 at $z = 4.1$. This is the brightest galaxy in the protocluster and inferred to be the dominant cluster galaxy in the process of formation. Because the equivalent width of Ly α is large ($\sim 500 \text{ \AA}$), the r band is dominated by Ly α . Arrows indicate the positions⁸ of the radio core (C) and the northern hotspot (H). The Ly α emission is elongated in the direction of the radio emission and the large-scale Ly α halo⁷ with a projected linear size of $\sim 15 \text{ kpc}$ (assuming $H_0 = 65 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$). **b**, Images of five spectroscopically confirmed Ly α emitters in the protocluster. Listed below each galaxy are its spectroscopic redshift⁷, the magnitude of the observed Lyman break, and the i-band magnitude. Two of the Ly α emitters are clumpy, as expected from young galaxies.

intervening intergalactic medium. Here we report multicolour imaging of the most distant candidate⁷⁻⁹ protocluster, TN J1338-1942 at a redshift $z \approx 4.1$. We find a large number of objects with the characteristic colours of galaxies at that redshift, and we show that this excess is concentrated around the targeted dominant radio galaxy. Our data therefore indicate that TN J1338-1942 is indeed the most distant cluster progenitor of a rich local cluster, and that galaxy clusters began forming when the Universe was only ten per cent of its present age.

There is increasing evidence that structures of galaxies existed in the early Universe, but the detection of protoclusters at redshifts $z > 1$ using conventional optical and X-ray techniques is difficult¹⁰⁻¹². Some of us have developed an efficient method for pinpointing distant protoclusters. The technique is based on the hypothesis that the most powerful known high-redshift radio galaxies are frequently associated with massive forming galaxies¹³⁻¹⁶ in protoclusters⁵. As a first step towards testing this hypothesis, we recently conducted a large programme with the Very Large Telescope (VLT) of the European Southern Observatory to search for galaxy overdensities associated with protoclusters around luminous high-redshift radio galaxies. Deep narrow- and broad-band imaging was used to locate candidate galaxies having bright Ly α emission, and follow-up spectra have confirmed that most of these candidates have similar redshifts to the high-redshift radio galaxies. All five targets studied with the VLT to sufficient depth have >20 spectroscopically confirmed Ly α and/or H α companion galaxies, associated with

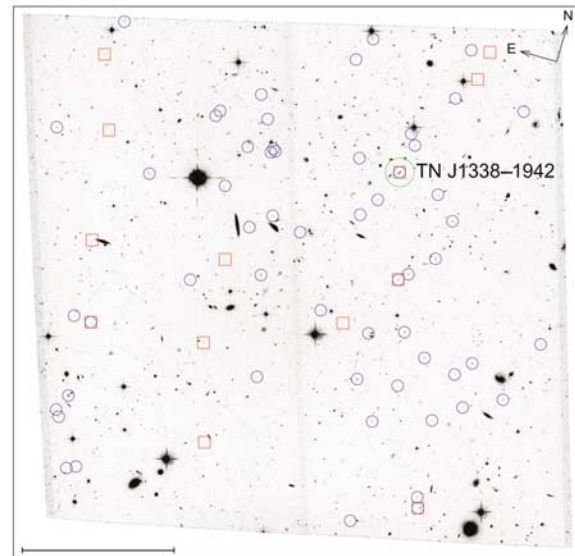


Figure 2 The spatial distribution of g-band dropout objects. Superimposed on the combined $3.4' \times 3.4'$ ACS greyscale image are the locations of g-band dropout objects (blue circles), selected to have colours and magnitudes of $(g-r) \geq 1.5$, $(g-r) \geq (r-i) + 1.1$, $(r-i) \leq 1$ and $i < 27$. In addition, objects were required to have a SExtractor²⁷ stellarity parameter of less than 0.85 to ensure that the sample was not contaminated by stars. These criteria filter galaxies having Lyman breaks at $z \approx 4$, thereby providing a sample of protocluster Lyman-break galaxy candidates. We detected 30 g-dropout objects in the field around TN J1338-1942 with $i_{775} < 26$, and 56 with $i_{775} < 27$. The number of g-band dropout objects is anomalously large, and their distribution is concentrated within a circular region of $\sim 1'$ in radius that includes the radio galaxy TN J1338-1942 (large green circle). Also shown are the positions of the spectroscopically confirmed Ly α emitters (red squares). Because the selection criteria were optimized to detect Lyman-break galaxies, some of the Ly α emitters did not fall into the formal sample of Lyman-break galaxies. The measured excess and its spatial clustering are evidence that a substantial fraction of the g-band dropout objects are Lyman-break galaxies associated with the protocluster. (See Fig. 1 legend for further details about the observations and the subsequent analysis.) Scale bar, $1''$.

galaxy overdensities^{6,7}. Their formal velocity dispersions are a few hundred km s⁻¹, but there was not enough time since the Big Bang for them to have become virialized. The scale sizes of the structures inferred from their spatial boundaries are ~3–5 Mpc. Assuming that the overdensities are due to a single structure, the masses derived from the observed structure sizes and overdensities are comparable to those of clusters of galaxies in the local Universe⁷. These observations led us to hypothesize that the overdensities of Ly α galaxies around radio sources are due to the fact that they are in protoclusters.

Galaxies that emit strong Ly α comprise only a small fraction of distant galaxies, and are biased towards non-dusty objects and galaxies that are undergoing the most vigorous star formation. Only about 25% of $z \approx 3$ galaxies have Ly α equivalent widths detectable by our VLT narrow-band imaging searches^{17,18}. If the overdensities of Ly α galaxies are located in protoclusters, additional galaxy populations should be present and detectable on the basis of characteristic continuum features in the galaxy spectra. The most important of these features is the sharp 'Lyman break' blueward of Ly α , caused by the absorption of the galaxy continuum radiation by neutral hydrogen clouds along the line of sight. Searching for Lyman-break galaxies is a powerful technique for finding high-redshift galaxies^{11,19,20}.

Because of its high spatial resolution, large field of view and excellent sensitivity, the Advanced Camera for Surveys²¹ (ACS) on the Hubble Space Telescope is uniquely suited for studying the morphologies of galaxies in the protoclusters and for finding additional galaxies on the basis of the Lyman-break features in their spectra. We therefore used the ACS to observe the most distant of our VLT protoclusters, TN J1338–1942 (ref. 7) at $z = 4.1$. This is a structure with 21 spectroscopically confirmed Ly α emitters and a rest-frame velocity dispersion of 325 km s⁻¹. Images were taken through three 'Sloan' filters—g band centred at 4,750 Å, r band centred near 6,250 Å and i band centred near 7,750 Å. These filters were chosen so that their wavelength responses bracketed redshifted Ly α at 6,214 Å and were sensitive to the Lyman-break feature blueward of Ly α .

A 3.4' $\times$ 3.4' field was observed, with the radio galaxy located ~1' from the image centre. Besides the radio galaxy, this field covered 12 of the 21 known Ly α emitting galaxies in the candidate protocluster. All 12 objects were detected in both r band and i band, with i-band magnitudes ranging from 25 to 28, compared with 23.3 ± 0.03 for the radio galaxy. As illustrated in Fig. 1, these objects were either absent or substantially attenuated in the g band, and their g–r colours are generally consistent with predicted values of Lyman breaks²². Half of the objects are extended in i_{775} , and three of these are resolved into two distinct knots of continuum emission, suggestive of merging.

We next used the Lyman-break technique to search for a population of Lyman-break galaxies in the protocluster that do not emit strong Ly α and would therefore have been undetectable in our VLT observations. Evidence for the existence of such a population was sought by analysing the number and spatial distribution of 'g-band dropout' objects—that is, objects whose colours are consistent with Lyman breaks in their spectra at the redshift of the protocluster. To investigate whether there is a statistically significant excess of such g-band dropout objects, we estimated the surface density and cosmic variance of g-band dropouts in a typical ACS field observed with the same filters and to the same depth as TN J1338–1942. We did this by cloning²³ B₄₃₅-band dropouts in 15 different pointings from the southern field of the Great Observatories Origins Deep Survey (GOODS)²⁴. Results indicate that the number of g-band dropouts in our field is a factor of 2.5 times higher than the average number found in a random GOODS field. Taking account of the typical cosmic variance²⁵ in the distribution of $z \approx 4$ Lyman-break galaxies, this is a 3 σ excess on the assumption that the distribution function is gaussian. Further evidence that a substantial fraction of

these g-dropout objects are Lyman-break galaxies associated with the protocluster is provided by the strong concentration of the g-band dropouts in a cluster-sized region around the radio galaxy. This is illustrated in Fig. 2. More than half of the g-band dropouts are located in a region of ~1' in radius (corresponding to a diameter of ~1 Mpc at $z = 4.1$). The number of g-band dropouts in this region is a factor of 5 times the average number encountered in similarly sized regions that are randomly drawn from the GOODS survey. This is a 5 σ excess, indicating that the number of g-band dropouts in our field is anomalously high at greater than the 99% confidence level. The spatial non-uniformity of g-band dropout objects in our field becomes even more pronounced when fainter objects down to a magnitude of $i = 27$ are included (Fig. 2).

Are there alternative explanations for the observed excess of g-band dropout objects other than a population of Lyman-break galaxies at $z \approx 4.1$? An object with a Balmer break at $z \approx 0.5$ could also be observed as a g-band dropout object. However, a population of such $z \approx 0.5$ objects would also be present in the GOODS comparison sample. Although the existence of an intervening structure of Balmer-break galaxies at $z \approx 0.5$ cannot be completely ruled out, its coincidence in location with the $z \approx 4.1$ structure of Ly α galaxies and the faintness and small sizes of the observed objects make this possibility highly unlikely.

The spatial coincidence of the excess in g-band dropout objects with the previously detected overdensity of Ly α emitters around a forming massive galaxy is strong evidence that we are observing a new population of Lyman-break galaxies in a protocluster. This would mean that TN J1338–1942, at $z \approx 4.1$, is indeed the most distant known protocluster, and that distant luminous radio galaxies pinpoint the progenitors of nearby rich clusters. Such protoclusters provide an opportunity to study the development of galaxies and clusters in the early Universe. They provide samples of different galaxy populations at the same distance, whose morphologies and spectral energy distributions could be used to disentangle the evolution and star formation history of different types of galaxies. The topological information that could be derived by mapping the shapes and sizes of such protoclusters over larger areas could answer the question of whether the first protoclusters in the early Universe formed in sheets or filaments. $\square$

Received 15 July; accepted 9 October 2003; doi:10.1038/nature02125.

- Kaiser, N. On the spatial correlation function of Abell clusters. *Astrophys. J.* **284**, L9–L12 (1984).
- White, S. D. M. & Rees, M. J. Core condensation in heavy halos—A two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183**, 341–358 (1978).
- Baugh, C. M., Cole, S., Frenk, C. S. & Lacey, C. G. The epoch of galaxy formation. *Astrophys. J.* **498**, 504–521 (1998).
- Bahcall, N. A. & Fan, X. The most massive distant clusters: Determining Ω and σ_8 . *Astrophys. J.* **504**, 1–6 (1998).
- Miley, G. in *Extrasolar Planets to Cosmology: The VLT Opening Symposium* (eds Bergeron, J. & Renzini, A.) 32–42 (Springer, Berlin, 2000).
- Pentericci, L. *et al.* A search for clusters at high redshift. II. A proto cluster around a radio galaxy at $z = 2.16$. *Astron. Astrophys.* **361**, L25–L28 (2000).
- Venemans, B. P. *et al.* The most distant structure of galaxies known: A protocluster at $z = 4.1$. *Astrophys. J.* **569**, L11–L14 (2002).
- De Breuck, C. *et al.* VLT spectroscopy of the $z = 4.11$ radio galaxy TN J1338–1942. *Astron. Astrophys.* **352**, L51–L56 (1999).
- De Breuck, C., van Breugel, W., Röttgering, H. J. A. & Miley, G. A sample of 669 ultra steep spectrum radio sources to find high redshift radio galaxies. *Astron. Astrophys. Suppl.* **143**, 303–333 (2000).
- Rosati, P. *et al.* An X-ray-selected galaxy cluster at $z = 1.26$. *Astron. J.* **118**, 76–85 (1999).
- Steidel, C. C. *et al.* A large structure of galaxies at redshift $z \sim 3$ and its cosmological implications. *Astrophys. J.* **492**, 428–438 (1998).
- Shimasaku, K. *et al.* Subaru deep survey IV: Discovery of a large-scale structure at redshift $z \sim 5$. *Astrophys. J.* **586**, L111–L114 (2003).
- De Breuck, C. *et al.* Optical and near-infrared imaging of ultra-steep-spectrum radio sources: The K-z diagram of radio-selected and optically selected galaxies. *Astron. J.* **123**, 637–677 (2002).
- Dey, A., van Breugel, W., Vacca, W. D. & Antonucci, R. Triggered star formation in a massive galaxy at $z = 3.8$: 4C 41.17. *Astrophys. J.* **490**, 698–709 (1997).
- Pentericci, L. *et al.* HST images and properties of the most distant radio galaxies. *Astrophys. J.* **504**, 139–146 (1999).
- van Ojik, R. *Gas in Distant Radio Galaxies: Probing the Early Universe* Thesis, Leiden Univ. (1995).
- Steidel, C. C. *et al.* Ly α imaging of a proto-cluster region at $\langle z \rangle = 3.09$. *Astrophys. J.* **532**, 170–182 (2000).
- Shapley, A. E., Steidel, C. C., Pettini, M. & Adelberger, K. Rest-frame ultraviolet spectra of $z \sim 3$ Lyman break galaxies. *Astrophys. J.* **588**, 65–89 (2003).

19. Steidel, C. C., Gialalisco, M., Pettini, M., Dickinson, M. & Adelberger, K. Spectroscopic confirmation of a population of normal star-forming galaxies at redshifts $z > 3$. *Astrophys. J.* **462**, L1–L7 (1999).
20. Steidel, C. C., Adelberger, K., Gialalisco, M., Dickinson, M. & Pettini, M. Lyman-break galaxies at $z > 4$ and the evolution of the ultraviolet luminosity density at high redshift. *Astrophys. J.* **519**, 1–17 (1999).
21. Ford, H. C. *et al.* Advanced camera for the Hubble Space Telescope. *Proc. SPIE* **3356**, 234–248 (1998).
22. Madau, P. Radiative transfer in a clumpy universe: The colors of high-redshift galaxies. *Astrophys. J.* **441**, 18–27 (1995).
23. Bouwens, R. J., Broadhurst, T. & Illingworth, G. Cloning dropouts: Implications for galaxy evolution at high redshift. *Astrophys. J.* **593**, 640–660 (2003).
24. Gialalisco, M. *et al.* The Great Observatories Origins Deep Survey. *Astrophys. J. Lett.* (special issue) (in the press).
25. Somerville, R. S. *et al.* Cosmic variance in the Great Observatories Origins Deep Survey. *Astrophys. J. Lett.* (special issue) (in the press); preprint at (<http://xxx.lanl.gov/astro-ph/0309071>) (2003).
26. Blakeslee, J. P., Anderson, K. R., Meurer, G. R., Benitez, N. & Magee, D. An automatic image reduction pipeline for the Advanced Camera for Surveys. *ASP Conf. Ser.* **295**, 257–260 (2003).
27. Bertin, E. & Arnouts, S. SExtractor: Software for source extraction. *Astron. Astrophys.* **117**, 393–404 (1996).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.K.M. (miley@strw.leidenuniv.nl).

Coherent spin manipulation without magnetic fields in strained semiconductors

Y. Kato, R. C. Myers, A. C. Gossard & D. D. Awschalom

Center for Spintronics and Quantum Computation, University of California, Santa Barbara, California 93106, USA

A consequence of relativity is that in the presence of an electric field, the spin and momentum states of an electron can be coupled; this is known as spin–orbit coupling. Such an interaction opens a pathway to the manipulation of electron spins within non-magnetic semiconductors, in the absence of applied magnetic fields. This interaction has implications for spin-based quantum information processing¹ and spintronics^{2,3}, forming the basis of various device proposals^{4–8}. For example, the concept of spin field-effect transistors^{4,5} is based on spin precession due to the spin–orbit coupling. Most studies, however, focus on non-spin-selective electrical measurements in quantum structures. Here we report the direct measurement of coherent electron spin precession in zero magnetic field as the electrons drift in response to an applied electric field. We use ultrafast optical techniques to spatiotemporally resolve spin dynamics in strained gallium arsenide and indium gallium arsenide epitaxial layers. Unexpectedly, we observe spin splitting in these simple structures arising from strain in the semiconductor films. The observed effect provides a flexible approach for enabling electrical control over electron spins using strain engineering. Moreover, we exploit this strain-induced field to electrically drive spin resonance with Rabi frequencies of up to ~ 30 MHz.

Even in the absence of a magnetic field, the conduction band of a semiconductor lacking an inversion centre has spin splitting away from the zone centre, owing to spin–orbit coupling. For example, the zinc blende crystal structure of GaAs has bulk inversion asymmetry (BIA), while heterostructures can introduce structural inversion asymmetry (SIA). BIA gives rise to the Dresselhaus spin splitting⁹, which has a quantization axis perpendicular to the electron wavevector $\mathbf{k}$, and changes its orientation with respect to $\mathbf{k}$ as the direction of $\mathbf{k}$ changes between the crystal directions $[110]$ and $[1\bar{1}0]$. Meanwhile, SIA gives rise to the Rashba hamiltonian¹⁰

($H_R = \alpha(\mathbf{k} \times \hat{\mathbf{n}}) \cdot \boldsymbol{\sigma}$, where α is the Rashba coefficient and $\boldsymbol{\sigma}$ is the Pauli operator) and introduces an effective magnetic field perpendicular to $\mathbf{k}$ and the symmetry-breaking axis $\hat{\mathbf{n}}$, but its relative orientation to $\mathbf{k}$ remains constant in the plane normal to $\hat{\mathbf{n}}$. For example, two-dimensional electron gases in semiconductor heterostructures have zero-magnetic-field splitting as a consequence of these mechanisms^{11,12}, and have been studied through the beatings in the Shubnikov de Haas oscillations^{13,14}, analysis of the antilocalization peak in magnetoresistance measurements^{15,16}, and spin-flip Raman scattering¹⁷. Gate-voltage control of the Rashba interaction has also been demonstrated in these structures¹⁸. Additionally, it has been predicted that these interactions may cause spin precession in the diffusive transport regime¹⁹. Although electrons scatter between many states with different $\mathbf{k}$, the application of an electric field should result in a non-zero average of the effective magnetic field that can coherently rotate the spins.

Here we report a similar spin splitting in bulk semiconductors originating from strain. We observe coherent electron spin precession angles exceeding 3π in zero magnetic field as the spins drift distances of greater than $60 \mu\text{m}$ in ~ 13 ns. Samples grown by molecular beam epitaxy²⁰ consist of a $2\text{-}\mu\text{m}$ n-GaAs layer ($n = 3 \times 10^{16} \text{ cm}^{-3}$) acting as a spin probe layer at the surface, and a $2\text{-}\mu\text{m}$ film of $\text{Al}_{0.4}\text{Ga}_{0.6}\text{As}$ underneath serving as a stressor/etch-stop layer. The semi-insulating GaAs (001) substrate is removed by chemical etching in order to form a rectangular membrane $\sim 100 \mu\text{m}$ by $\sim 300 \mu\text{m}$. The processed membrane has curvature, possibly owing to the larger lattice constant or the oxidation of the $\text{Al}_{0.4}\text{Ga}_{0.6}\text{As}$ layer, thereby straining the n-GaAs film. Ni/GeAu ohmic contacts are formed on the surface in order to apply an in-plane electric field E along $[1\bar{1}0]$.

Time- and spatially resolved Faraday rotation (FR) spectroscopy^{21,22} is used to probe the electron spin dynamics. In this pump–probe measurement, the FR of the probe beam measures the total magnetization of the optically injected electron spins in the growth direction z . The laser beams are normally incident on the sample and focused to spots with full-width at half-maximum of $14 \mu\text{m}$, whose relative distance d along the direction of E can be controlled with a stepper-motor-driven mirror, enabling spatial scanning in one dimension. All the measurements are taken at temperature $T = 5$ K in a magneto-optical cryostat, and the external

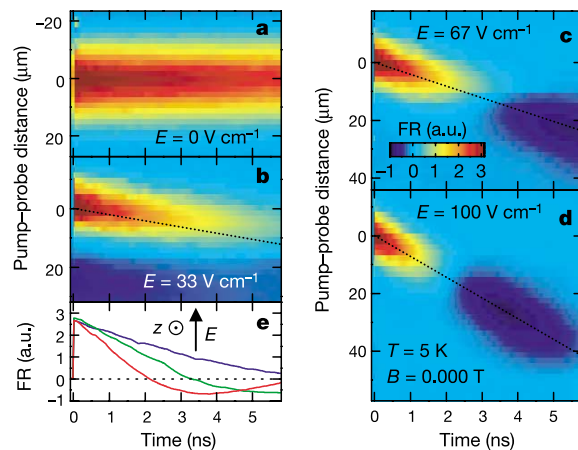


Figure 1 Spatiotemporal evolution of a spin packet at zero magnetic field. **a–d**, FR as a function of temporal and spatial separation between pump and probe pulses for $E = 0$, 33, 67, 100 V cm^{-1} , respectively. Data in **a** have been multiplied by 0.67 for clarity, and all panels share the colour scale shown in **c**. The dotted lines have slopes determined by the spin drift velocity. The slight negative signal in the left bottom corner of **b** is due to an older spin packet created by a previous pulse. **e**, Line sections along dotted lines in **b** (blue), **c** (green) and **d** (red). Inset shows the geometry. Zero external magnetic field was calibrated by the resonant spin amplification peak at $E = 0 \text{ V cm}^{-1}$.

magnetic field B_{ext} is monitored with an *in situ* Hall sensor.

Figure 1a–d shows the spatiotemporal evolution of a coherent electron spin packet with $B_{\text{ext}} = 0.0000 \pm 0.0002$ T, under various E . The spin packet drifts and coherently precesses as time evolves, and as previously observed²² its velocity increases with increasing E . Line sections along the motion of the packet (Fig. 1e) show, however, that the precession frequency itself is also increasing with E . The observed internal effective magnetic field B_{int} is accurately characterized by analysing the B_{ext} dependence of the FR at a fixed delay²². In the Voigt geometry with $E = 0$ V cm⁻¹, the FR signal can be described by $S_0 \exp(-\Delta t/T_2^*) \cos(g\mu_B B_{\text{ext}} \Delta t/\hbar)$, where S_0 is the initial amplitude, Δt the temporal delay between the pump and probe pulses, T_2^* the inhomogeneous transverse electron spin lifetime, g the electron g -factor, μ_B the Bohr magneton, and $\hbar$ the Planck constant. The signal is periodic as a function of B_{ext} and symmetric around $B_{\text{ext}} = 0$ T at a fixed $\Delta t = 13.1$ ns (top curve of Fig. 2a). The sharp peaks are due to a summation of contributions from successive pulses at the laser repetition rate²³. When a large enough E is applied, the spin packets are spatially separated, and it is possible to select the contribution from the most recent pulse. In this case, one expects a symmetric cosine function, but the presence

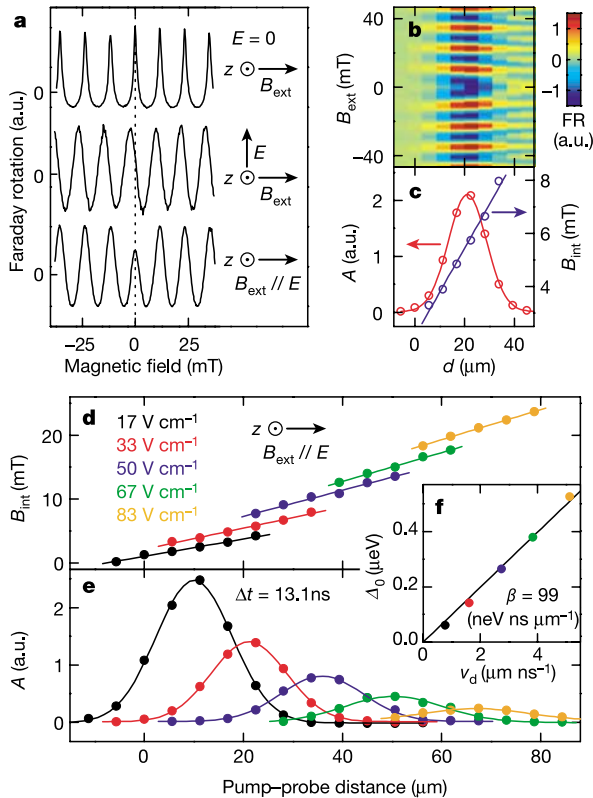


Figure 2 Characterization of internal field. **a**, B_{ext} dependence of FR at $\Delta t = 13.1$ ns for $E = 0$ V cm⁻¹ at $d = 0.0$ μm (top trace), $E = 17$ V cm⁻¹ in $E \perp B_{\text{ext}}$ geometry at $d = 7.9$ μm (middle trace), and $E = 17$ V cm⁻¹ in $E//B_{\text{ext}}$ geometry at $d = 5.6$ μm (bottom trace). The curves are offset and scaled for clarity, and the insets show the geometries. **b**, FR at $\Delta t = 13.1$ ns as a function of d and B_{ext} with $E = 33$ V cm⁻¹. **c**, Fitting parameters A (red circles) and B_{int} (blue circles) for the data in **b** as a function of d , along with fits (lines) as explained in the text. **d** and **e**, B_{int} and A , respectively, as a function of d for $E = 17$ (black), 33 (red), 50 (blue), 67 (green), 83 V cm⁻¹ (orange) in $E//B_{\text{ext}}$ geometry. Solid symbols are data, and lines are fits, as described in the text. Spin diffusion constant D can be obtained from the width of the spin packet or the slope of B_{int} with respect to position, taking into account the signal convolution over the laser spot size. D increases from 1.0 $\mu\text{m}^2 \text{ns}^{-1}$ at $E = 17$ V cm⁻¹ to 3.3 $\mu\text{m}^2 \text{ns}^{-1}$ at $E = 83$ V cm⁻¹; the numbers obtained by these two methods agree within 0.3 $\mu\text{m}^2 \text{ns}^{-1}$, and show similar E dependence. **f**, Δ_0 as a function of v_d . The solid line is a linear fit to data.

of B_{int} modifies the line shape.

When E is applied perpendicular to B_{ext} , the curve is shifted in the B_{ext} axis (Fig. 2a, middle curve), indicating that B_{int} is along B_{ext} , that is, perpendicular to both z and E . The sign of B_{int} can be determined from the direction of the shift, and it changes when the direction of E is reversed (Supplementary Figure). For the case of E and B_{ext} being parallel (Fig. 2a, bottom curve), assuming that B_{int} is independent of B_{ext} , the data can be fitted to $A \cos(g\mu_B \sqrt{B_{\text{ext}}^2 + B_{\text{int}}^2} \Delta t/\hbar)$, as B_{int} is perpendicular to B_{ext} in this geometry²². FR is measured for various d at $E = 33$ V cm⁻¹ (Fig. 2b) and the parameters A and B_{int} are plotted in Fig. 2c as a function of d . The d dependence of B_{int} can be explained by the distribution of spin drift velocity v_d along the spin packet owing to spin diffusion²⁴. In order to accurately evaluate B_{int} , we fit A with a gaussian function to determine the centre of the spin packet and subsequently use a linear fit to B_{int} versus d in order to extract the value at that position. The procedure is repeated for different E (Fig. 2d and e), and the spin splitting $\Delta_0 = g\mu_B B_{\text{int}}$ at the centre of the spin packet is plotted in Fig. 2f as a function of v_d . We introduce a phenomenological linear relation $\Delta_0 = \beta v_d$ and use the coefficient β to characterize the observed effect.

Although one possible origin of B_{int} is a current-induced magnetic field, it seems unlikely, because the sign of such magnetic fields changes along z whereas B_{int} has a definite sign. Moreover, a calculation assuming uniform current density yields a value three orders of magnitude smaller than the observed B_{int} . Surprisingly, we find that strain plays a critical role in generating B_{int} . Additional samples from the same wafer were prepared without the substrate removal process to avoid strain. Similar measurements using Kerr rotation spectroscopy were performed, and β is suppressed by an order of magnitude for both E along the $[110]$ and $[1\bar{1}0]$ directions (Fig. 3a).

There are at least two possible mechanisms for strain-induced zero-magnetic-field spin splitting. Uniform strain can modify the magnitude of the BIA term, whereas a strain gradient can break the inversion symmetry, thus introducing SIA. In the latter case, the axis of asymmetry needs to be in the z direction in order to produce B_{int} in the plane of the sample for in-plane k . We investigate the role of the two mechanisms by introducing strain in lattice-mismatched heterostructures. A series of samples having $\text{In}_{0.07}\text{Ga}_{0.93}\text{As}$ layers on a GaAs substrate were prepared, and results from two representative samples are shown in Fig. 3b. The first sample has 200 nm

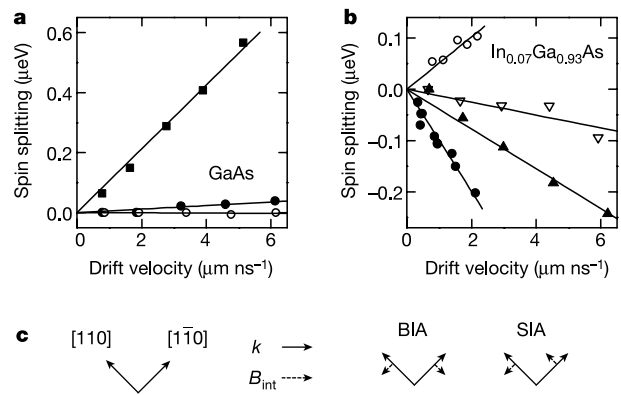


Figure 3 Strain-induced nature of the internal field. **a**, Δ_0 as a function of v_d for GaAs film samples processed into a membrane (filled squares) and left attached to the substrate ($E//[110]$, open circles, and $E//[1\bar{1}0]$, filled circles). **b**, Δ_0 as a function of v_d for 200-nm (circles) and 2,000-nm (triangles) $\text{In}_{0.07}\text{Ga}_{0.93}\text{As}$ layers. $E//[110]$, open symbols, and $E//[1\bar{1}0]$, filled symbols. The solid lines are linear fits to data, giving $\beta_{\text{BIA}} = 75$ neV ns μm^{-1} and $\beta_{\text{SIA}} = -24$ neV ns μm^{-1} for the 200-nm sample, and $\beta_{\text{BIA}} = 13$ neV ns μm^{-1} and $\beta_{\text{SIA}} = -26$ neV ns μm^{-1} for the 2,000-nm sample. **c**, Diagram showing the orientations of the BIA and SIA components of B_{int} .

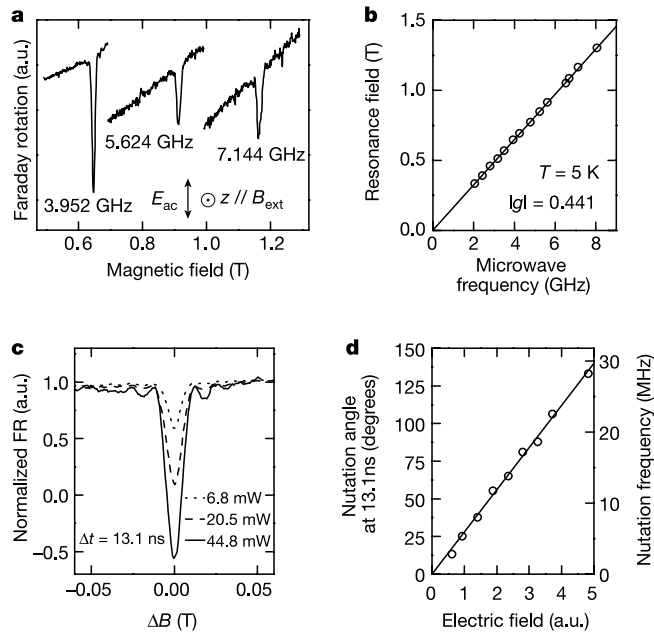


Figure 4 Electrically driven spin resonance using strain-induced field. **a**, B_{ext} dependence of FR in Faraday geometry at $\Delta t = 13.1$ ns with $\nu = 3.952$, 5.624 and 7.144 GHz. The sharp peaks are due to electrically driven spin resonance. The inset shows the geometry of the experiment. **b**, Position of the resonance as a function of ν . **c**, Fine scans of FR normalized to the baseline value near the resonance at $\Delta t = 13.1$ ns with microwave powers of 6.8 (dotted line), 20.5 (dashed line) and 44.8 mW (solid line) applied at the top of the cryostat for $\nu = 2.052$ GHz. The width of the peak, corresponding to $T_2^* \approx 6$ ns, does not noticeably increase with increasing microwave power. The small oscillations in the tails of the dip result from integration of the transverse field. In the rotating frame, detuning causes the transverse field to rotate and cancel out most of its effect when it has rotated an integer number of times at that Δt . **d**, Nutation angle at $\Delta t = 13.1$ ns and nutation frequency as a function of E_{ac} for $\nu = 2.052$ GHz. The data point for the highest electric field was taken with microwave power of 44.8 mW applied at the top of the cryostat.

of $\text{n-In}_{0.07}\text{Ga}_{0.93}\text{As}$, whereas the second sample has 2,000 nm of $\text{In}_{0.07}\text{Ga}_{0.93}\text{As}$ whose top 500 nm is n-doped with the rest remaining undoped. Strain in the out-of-plane direction is determined by X-ray diffraction, and is +0.46% for the 200-nm sample and +0.14% for the 2,000-nm sample assuming an equilibrium lattice constant of 5.682 Å.

As the BIA component changes sign between [110] and $[1\bar{1}0]$ directions while the SIA component keeps its orientation (Fig. 3c), the two contributions can be identified by considering $\beta_{\text{BIA}} = (\beta_{[110]} - \beta_{[1\bar{1}0]})/2$ and $\beta_{\text{SIA}} = (\beta_{[110]} + \beta_{[1\bar{1}0]})/2$. We performed measurements in the two crystal directions in order to determine β_{BIA} and β_{SIA} . The 200-nm sample shows a larger β_{BIA} , suggesting that the BIA term is enhanced owing to strain. In contrast, β_{SIA} ranges from -26 to 69 $\text{neV ns } \mu\text{m}^{-1}$ throughout the samples investigated, and does not vary systematically with strain (Supplementary Table). Possible sources of SIA are unintentional drift in the In concentration or strain gradients. Compositional drift is ruled out, as samples with linearly graded In concentration by digital growth²⁰ (0–5% over 500 nm) have β_{SIA} values of the same magnitude. Strain gradient, if any, is expected to change with InGaAs thickness, but samples with total InGaAs thickness of 500, 1,000 and 1,500 nm show β_{SIA} values that agree within ± 13 $\text{neV ns } \mu\text{m}^{-1}$. This is too small a variation to explain the larger deviations in other samples. Anisotropic in-plane strain may also cause spin splitting of the same symmetry²⁵, but no systematic correlation is observed. Further studies are needed to clarify the origin of the SIA contribution.

Finally, we show that B_{int} can be effectively used to electrically drive spin resonance^{26,27}. We apply an electric field E_{ac} at a microwave frequency ν using a phase-locked oscillator²⁸ to the GaAs membrane sample (inset, Fig. 4a). The FR signal is measured at $d = 0$ μm and $\Delta t = 13.1$ ns as B_{ext} is swept. We see a dip in the signal at resonant field (Fig. 4a), and the dip position shifts linearly with ν (Fig. 4b). A linear fit gives $|g| = 0.441$, consistent with other FR measurements on this sample. In Fig. 4c, it can be seen that the signal drops below zero for the highest microwave power, which shows that the process is a coherent nutation rather than incoherent depolarization of electron spins. A nutation angle can be estimated by taking the arccosine of the ratio of signal minimum and the baseline value, and is plotted as a function of E_{ac} in Fig. 4d. The nutation frequency increases linearly with E_{ac} as expected, reaching ~ 30 MHz.

The strain-induced spin splitting enables electrical control of electron spins in the absence of magnetic fields, providing a distinct pathway to solid-state spin manipulation in contrast to other recently developed techniques^{28–30}. By adding functionality using strain, simple epilayers, rather than sophisticated quantum structures, could be used for spintronics applications. Advances in strain engineering, exploited in conventional electronic devices, could be redirected to provide new opportunities for quantum computation schemes and spintronic devices^{4–8}. For example, micromechanical structures could be used to control strain, and thereby spins, through piezoelectric modulation. □

Methods

Time-resolved Faraday and Kerr rotation spectroscopy

A mode-locked Ti:sapphire laser produces ~ 150 -fs pulses with a repetition frequency of 76 MHz, which are split into a circularly polarized pump beam (20 μW for GaAs membrane sample, 400–950 μW otherwise) and a linearly polarized probe beam (15–35 μW). The pump beam creates a spin population in the conduction band at time $t = 0$, while the probe beam transmits through (Faraday rotation spectroscopy) or reflects off (Kerr rotation spectroscopy) the sample at $t = \Delta t$. The polarization of the probe beam rotates by an amount proportional to the spin component along its propagation direction²¹. The pump polarization is modulated between right and left circular polarization with a photoelastic modulator at frequency f_1 while the probe beam is chopped at frequency f_2 , and the rotation angle is lock-in-detected with a balanced photodiode bridge at frequency $f_1 \pm f_2$. The laser wavelength is tuned to 816–820 nm for GaAs samples and 866–902 nm for $\text{In}_{0.07}\text{Ga}_{0.93}\text{As}$ samples.

Sample structures and strain characterization

The GaAs membrane sample has a 300- μm -long, 60- μm -wide rectangular mesa with a resistance of ~ 1 k Ω . Strain in this sample was estimated using optical interference fringes at room temperature. The highest point of the membrane was 700 ± 200 nm above the substrate, which would lead to an in-plane strain of $\sim 10^{-2}$. Smaller strain in this sample relative to the InGaAs samples suggests that additional mechanisms, such as strain gradient, are contributing to the spin splitting.

A series of $\text{In}_{0.07}\text{Ga}_{0.93}\text{As}$ layers were grown on GaAs substrates, and strains in these films were characterized by reciprocal space mapping around GaAs [115] and $[1\bar{1}\bar{5}]$ peaks with an X-ray diffractometer at room temperature. The samples were capped with undoped GaAs to avoid surface depletion. A sample with undoped $\text{In}_{0.13}\text{Ga}_{0.87}\text{As}$ capping showed no marked effect on β , indicating that capping does not play a significant role. Spins in the undoped region are not observed at $\Delta t = 13.1$ ns, as the spin lifetime in this region is < 1 ns. Some samples showed lateral inhomogeneity in β , possibly owing to inhomogeneity in strain relaxation.

Received 20 August; accepted 3 November 2003; doi:10.1038/nature02202.

- Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
- Awschalom, D. D., Loss, D. & Samarth, N. (eds) *Semiconductor Spintronics and Quantum Computation* (Springer, Berlin, 2002).
- Wolf, S. A. *et al.* Spintronics: a spin-based electronics vision for the future. *Science* **294**, 1488–1495 (2001).
- Datta, S. & Das, B. Electronic analog of the electro-optic modulator. *Appl. Phys. Lett.* **56**, 665–667 (1990).
- Schliemann, J., Egues, J. C. & Loss, D. Nonballistic spin-field-effect transistor. *Phys. Rev. Lett.* **90**, 146801 (2003).
- Voskoboynikov, A., Lin, S. S. & Lee, C. P. Spin-polarized electronic current in resonant tunneling heterostructures. *J. Appl. Phys.* **87**, 387–391 (2000).
- Koga, T., Nitta, J., Takayanagi, H. & Datta, S. Spin-filter device based on the Rashba effect using a nonmagnetic resonant tunneling diode. *Phys. Rev. Lett.* **88**, 126601 (2002).
- Kiselev, A. A. & Kim, K. W. T-shaped ballistic spin filter. *Appl. Phys. Lett.* **78**, 775–777 (2001).
- Dresselhaus, G. Spin-orbit coupling effects in zinc blende structures. *Phys. Rev.* **100**, 580–586 (1955).
- Bychkov, Y. A. & Rashba, E. I. Oscillatory effects and the magnetic susceptibility of carriers in inversion layers. *J. Phys. C* **17**, 6039–6045 (1984).

11. Pfeffer, P. Effect of inversion asymmetry on the conduction subbands in GaAs-Ga_{1-x}Al_xAs heterostructures. *Phys. Rev. B* **59**, 15902–15909 (1999).
12. Lommer, G., Malcher, F. & Rossler, U. Spin splitting in semiconductor heterostructures for B → 0. *Phys. Rev. Lett.* **60**, 728–731 (1988).
13. Das, B., Datta, S. & Reifenger, R. Zero-field spin splitting in a two-dimensional electron gas. *Phys. Rev. B* **41**, 8278–8287 (1990).
14. Luo, J., Munkata, H., Fang, F. & Stiles, P. J. Effects of inversion asymmetry on electron energy band structures in GaSb/InAs/GaSb quantum wells. *Phys. Rev. B* **41**, 7685–7693 (1990).
15. Dresselhaus, P. D., Papavassiliou, C. M. A., Wheeler, R. G. & Sacks, R. N. Observation of spin precession in GaAs inversion layers using antilocalization. *Phys. Rev. Lett.* **68**, 106–109 (1992).
16. Knap, W. *et al.* Weak antilocalization and spin precession in quantum wells. *Phys. Rev. B* **53**, 3912–3924 (1996).
17. Jusserand, B., Richards, D., Peric, H. & Etienne, B. Zero-magnetic-field spin splitting in the GaAs conduction band from Raman scattering on modulation-doped quantum wells. *Phys. Rev. Lett.* **69**, 848–851 (1992).
18. Nitta, J., Akazaki, T., Takayanagi, H. & Enoki, T. Gate control of spin-orbit interaction in an inverted In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As heterostructure. *Phys. Rev. Lett.* **78**, 1335–1338 (1997).
19. Kalevich, V. K. & Korenov, V. L. Effect of electric field on the optical orientation of 2D electrons. *JETP Lett.* **52**, 230–235 (1990).
20. Gossard, A. C. Growth of microstructures by molecular beam epitaxy. *IEEE J. Quant. Electron.* **QE-22**, 1649–1655 (1986).
21. Crooker, S. A., Awschalom, D. D., Baumberg, J. J., Flack, F. & Samarth, N. Optical spin resonance and transverse spin relaxation in magnetic semiconductor quantum wells. *Phys. Rev. B* **56**, 7574–7588 (1997).
22. Kikkawa, J. M. & Awschalom, D. D. Lateral drag of spin coherence in gallium arsenide. *Nature* **397**, 139–141 (1999).
23. Kikkawa, J. M. & Awschalom, D. D. Resonant spin amplification in n-type GaAs. *Phys. Rev. Lett.* **80**, 4313–4316 (1998).
24. Flatté, M. E. & Byers, J. M. Spin diffusion in semiconductors. *Phys. Rev. Lett.* **84**, 4220–4223 (2000).
25. La Rocca, G. C., Kim, N. & Rodriguez, S. Effect of uniaxial stress on the electron spin resonance in zinc-blende semiconductors. *Phys. Rev. B* **38**, 7595–7601 (1988).
26. Rashba, E. I. & Sheka, V. I. Combinational resonance of zonal electrons in crystals having a zinc blende lattice. *Sov. Phys. Solid State* **3**, 1257–1267 (1961).
27. Rashba, E. I. & Efros, A. L. Orbital mechanisms of electron-spin manipulation by an electric field. *Phys. Rev. Lett.* **91**, 126405 (2003).
28. Kato, Y. *et al.* Gigahertz electron spin manipulation using voltage-controlled g-tensor modulation. *Science* **299**, 1201–1204 (2003).
29. Salis, G. *et al.* Electrical control of spin coherence in semiconductor nanostructures. *Nature* **414**, 619–622 (2001).
30. Gupta, J. A., Knobel, R., Samarth, N. & Awschalom, D. D. Ultrafast manipulation of electron spin coherence. *Science* **292**, 2458–2461 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. M. Andrews, E. L. Hu, P. M. Petroff and J. S. Speck for discussions. This work was supported by the DARPA SPINS and QuIST programmes.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.D.A. (awsch@physics.ucsb.edu).

The interface between silicon and a high-*k* oxide

Clemens J. Först^{1,2}, Christopher R. Ashman¹, Karlheinz Schwarz² & Peter E. Blöchl¹

¹Clausthal University of Technology, Institute for Theoretical Physics, Leibnitzstrasse 10, D-38678 Clausthal-Zellerfeld, Germany

²Vienna University of Technology, Institute for Materials Chemistry, Getreidemarkt 9/165-TC, A-1060 Vienna, Austria

The ability of the semiconductor industry to continue scaling microelectronic devices to ever smaller dimensions (a trend known as Moore's Law¹) is limited by quantum mechanical effects: as the thickness of conventional silicon dioxide (SiO₂) gate insulators is reduced to just a few atomic layers, electrons can tunnel directly through the films. Continued device scaling will therefore probably require the replacement of the insulator with high-dielectric-constant (high-*k*) oxides², to increase its

thickness, thus preventing tunnelling currents while retaining the electronic properties of an ultrathin SiO₂ film. Ultimately, such insulators will require an atomically defined interface with silicon without an interfacial SiO₂ layer for optimal performance. Following the first reports of epitaxial growth of AO and ABO₃ compounds on silicon^{3–7}, the formation of an atomically abrupt crystalline interface between strontium titanate and silicon was demonstrated^{8–10}. However, the atomic structure proposed for this interface is questionable because it requires silicon atoms that have coordinations rarely found elsewhere in nature. Here we describe first-principles calculations of the formation of the interface between silicon and strontium titanate and its atomic structure. Our study shows that atomic control of the interfacial structure by altering the chemical environment can dramatically improve the electronic properties of the interface to meet technological requirements. The interface structure and its chemistry may provide guidance for the selection process of other high-*k* gate oxides and for controlling their growth.

Interfacing a new oxide with silicon is a major challenge. A gate oxide has to fulfil a number of requirements in addition to intrinsic properties such as the high dielectric constant and low defect concentrations. The gate oxide must also be chemically stable in contact with silicon, it must have a sufficiently large injection barrier, and it must not have interface states in the bandgap of silicon. Oxides related to silicon dioxide, used in today's transistors, uniquely fulfil these requirements owing to their strong bonds to silicon and their flexible bonding network. They fail only because of their low dielectric constant. Retaining the same beneficial properties for high-*k* oxides has turned out to be very difficult. The first high-*k* oxides introduced technologically are likely to be amorphous oxides or silicates of Hf and Zr with an interfacial SiO₂ layer. By around 2010, however, the projected miniaturization of transistors will mean that an interfacial SiO₂ layer will no longer be tolerable and oxides with an atomically well-defined interface with silicon will be required.

Before discussing the formation of the interface, we need to review the clean (001) surface of silicon and describe its changes due to Sr adsorption. On the unreconstructed silicon surface the atoms form a square array. Owing to a lack of upper bonding partners, each atom has two singly occupied dangling bonds pointing out of the surface. Pairs of silicon atoms dimerize, using up one dangling bond per atom to form the dimer bond. This is called the 'dimer row' reconstruction. A second rearrangement leads to the 'buckled dimer' reconstruction, in which one atom of each dimer lifts up and the other shifts down, resulting in a buckled dimer. This buckling causes both electrons to localize in the upper silicon atom of a dimer, whereas the other silicon atom with the empty dangling bond prefers a more planar arrangement.

By depositing various amounts of Sr atoms onto the Si(001) surface we explored the structural complexity of Sr adlayers¹¹. This enables us to attribute atomic structures to the periodicities observed experimentally^{10,12}. Here we summarize only the findings relevant for the present topic: initially each Sr atom donates two electrons into the empty dangling bonds of the surface. As Sr is added, the dimer buckling vanishes because both dangling bonds of a Si dimer become filled with electrons. Similar to Ba (refs 13, 14), Sr first occupies the trough between the dimer rows, in the centre of four dimers. At a coverage of half a monolayer all positions in the trough are occupied and each dimer dangling bond is filled with two electrons. This (2 × 1) structure is the only Sr-covered surface without surface states in the bandgap of silicon. Therefore, it is a suitable building block for an interface without states in the gap, as required for device applications. The finding of a sizeable bandgap also explains why this surface is fairly resistant to oxidation¹⁵.

Thus we found that the substrate surface can be chemically

saturated by half a monolayer of Sr. Such a surface is ‘isoelectronic’ to an H-terminated silicon surface. Hydrogen is known to be very effective at passivating silicon. In the following, we therefore refer to the Si surface covered with half a monolayer of Sr as the Sr-passivated substrate.

After having gained insight into the metal overlayers, we investigated the formation of an oxide layer. We start from the Sr-passivated substrate and simulate the deposition of one layer of SrO. During a heating cycle to 600 K this single oxide layer reconstructs significantly. However, after placing two or more layers of SrO or SrTiO₃ on top of the reconstructed SrO layer, the oxide layers crystallize into the perfect bulk structure. Thus we obtain an atomically abrupt interface between the silicon substrate and the high-*k* oxide. This interface structure, denoted A and shown in Fig. 1, corresponds to the Sr-passivated silicon surface matched to the nonpolar SrO layer of the oxide.

After the chemical saturation of the substrate surface with half a monolayer of Sr, the second-most-important property is the matching of the charge patterns of the oxide and the Sr-passivated substrate surfaces joined at the interface. Whereas the SiO₂/Si interface relies on strong covalent bonds across the interface and a flexible bond network of the oxide, the interface described here is

based on the chemical saturation of the silicon surface with an alkaline earth metal, so that a template for the deposition of a matching oxide is obtained.

In a device the interface is exposed to a number of chemical influences that affect the stability of the stack. The most critical question is the stability of the interface with respect to oxidation. Oxygen ions can diffuse out of the gate oxide to the interface.

To explore how the interface changes upon oxidation, we have added oxygen atoms to a wide range of different sites. Oxygen first attacks the surface silicon atoms at their vacant coordination sites. After introducing one monolayer of oxygen into the interface all these sites are consumed. The resulting structure, denoted B, is shown in Fig. 1. Additional oxygen atoms, up to a total oxygen content of 1.5 monolayers, insert into the dimer bonds. As explained below, structure B and the dimer-oxidized variant of structure B are the optimum choices for device applications.

Figure 2 illustrates the phase boundaries of the interfaces as a function of the oxygen chemical potential. The chemical potential is defined as the energy required to add a single oxygen atom to the system. It is the driving force for oxidation, which can be controlled externally, for example, by choosing the appropriate temperature

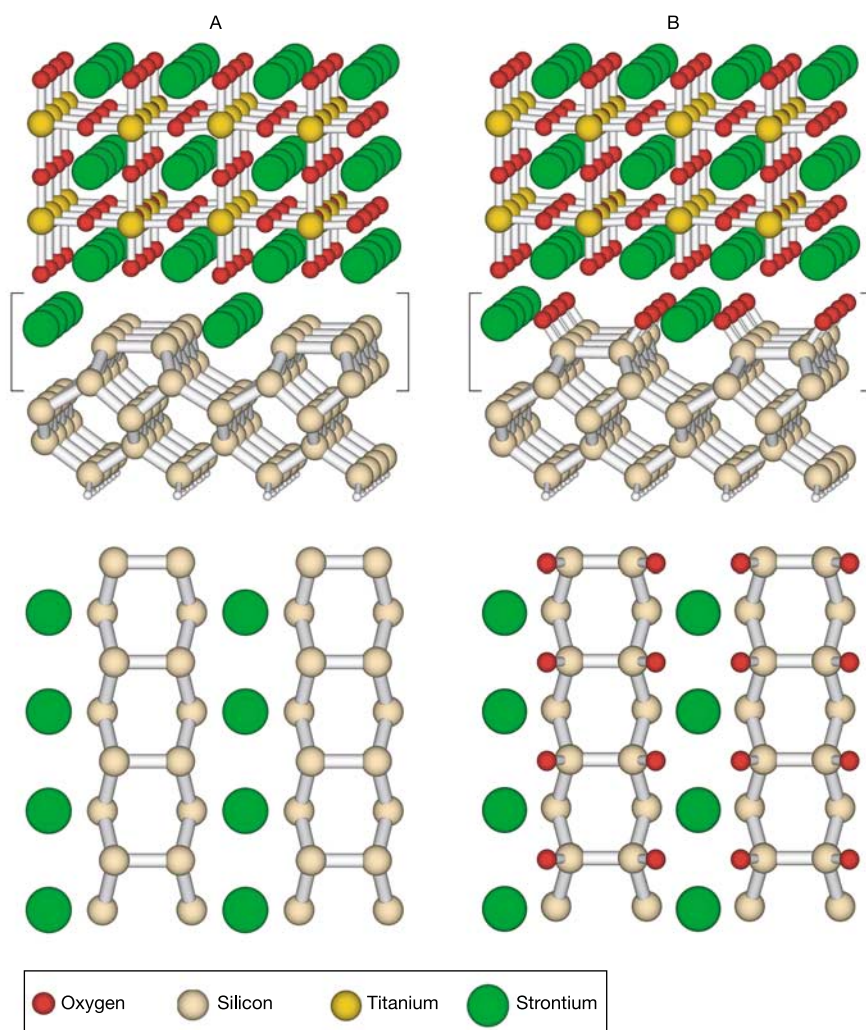


Figure 1 Atomic structures of the SrTiO₃/Si(001) interfaces: structure A (on left), unoxidized interface; structure B (on right), oxidized interface. Top, view slightly off the [110] direction of silicon, which is parallel to the [100] direction of the oxide. The topmost

layer corresponds to the oxide surface of our slab calculation. Bottom, view along the interface normal of the interface layer.

and oxygen partial pressure during growth. In thermal equilibrium the chemical potential is related to the partial pressure and temperature of the growth chamber. It should be noted, however, that the formation of the stack mostly involves non-equilibrium processes, so that the chemical potential at the interface lags behind the value reached in the growth chamber. The oxygen chemical potential needs to be sufficiently low to avoid formation of an interfacial SiO_2 layer but high enough to avoid a large oxygen vacancy concentration in the gate oxide, which may cause trap-assisted leakage currents.

Our results indicate that structures A and B can be formed selectively in the absence of an interfacial SiO_2 layer. The processing window—the stability region of interface B—extends from -0.24 to 0.05 eV and corresponds to a range of partial pressures of nearly three orders of magnitude at $1,000$ K. This region lies almost entirely below the coexistence line of Si and SiO_2 .

In addition, we find indications of sizeable thin-film effects, which delay the formation of an interfacial SiO_2 layer. This can be inferred from the fact that the oxidation of all dimer bonds requires a chemical potential of 0.19 eV, and subsurface oxidation starting from interfaces A or B requires energies in excess of 0.91 eV above the coexistence line of Si and SiO_2 .

The interface structure proposed in this work is quite different and much simpler than previously assumed: it was believed that an interfacial silicide^{8,9} or silicate^{16,17} layer must be formed. Such a layer is not present in the interface that emerged from our simulations. Nevertheless, our simulated $\text{SrTiO}_3/\text{Si}(001)$ interfaces reproduce the undisputed features of the Z-contrast images of ref. 8, such as the pattern of interfacial Sr atoms and the oxide-substrate registry. Thus the lateral alignment of the columns of Sr and Ti atoms relative to the Si substrate and the (2×1) periodicity can clearly be identified.

We also performed calculations on the interfaces proposed by McKee and co-workers^{8,9,10} and by Wang and co-workers^{16,17}. As also shown previously¹⁸, the former interface^{8,9} is metallic, which is detrimental for device applications. The same applies for the latter structural proposal^{16,17}, which reconstructs significantly upon relaxation.

On the basis of electron-count arguments, Robertson and Peacock recently proposed a structure¹⁸ that is related to our

dimer-oxidized variant of structure B. It differs in that the oxide starts with the TiO_2 layer instead of a SrO layer and it is derived from a $c(2 \times 2)$ dimer reconstruction of the silicon surface. If we modify our interface by terminating the oxide with a TiO_2 layer instead of a SrO layer, it is more stable than Robertson's proposal by 0.19 eV per (1×1) surface unit cell. Having an oxide terminated by SrO is, however, favourable compared to TiO_2 terminated oxides, because the TiO_2 layer and the substrate in direct contact are expected to react, as pointed out¹⁸. Conceptually similar interfaces with a TiO_2 interface layer have been investigated¹⁹, ruling out their use in devices on the basis of their electronic properties.

A critical parameter for gate stacks is the injection barrier, which is the offset between the conduction band edges of the silicon substrate and the oxide. It prevents electrons from entering the oxide conduction band, where they can cross the gate oxide. For device applications the injection barrier should be larger than 1 eV (ref. 3). There are indications^{20,21} that the injection barriers for most high- k oxides are too low.

Before we discuss our results on the band offsets we need to briefly touch upon the bandgap problem of density functional theory (DFT)^{22,23}: typically the one-particle energies obtained in these calculations underestimate the bandgap. Therefore, there is an uncertainty in our calculated injection barriers. Assuming that the error of the valence band edge is negligible, as required in exact DFT, and using the experimental bandgaps of silicon and SrTiO_3 (ref. 24), we anticipate that our calculations underestimate the injection barrier by 0.7 – 0.8 eV.

For interface A we obtain an injection barrier that is negative by 0.6 eV. Including our correction, we estimate the injection barrier to lie at 0.1 – 0.2 eV. The injection barrier lies below the technologically required minimum.

For interface B, however, we obtain a positive injection barrier of 0.5 eV. Adding the correction, our final estimate yields a positive injection barrier of 1.2 – 1.3 eV, which fulfils the criterion. The margin is sufficiently large that a reasonable error in the band-offset correction does not lead to an unacceptably low injection barrier. Most important is the ability to influence the injection barrier by carefully choosing the processing conditions.

Note that band offsets are frequently derived from the properties of the two bulk materials alone²¹, disregarding the interface structure and composition in detail. The interface between silicon and SrTiO_3 is an example where the band offset can be engineered by controlling the chemical environment. The change of the band offset due to oxidation is about 1.1 eV, and is thus sizeable. It results from a dipole created when the electrons are transferred from the filled dangling bonds of the surface silicon atoms to the oxygen atoms that attach to the vacant coordination sites.

The injection barrier of interface B is fairly insensitive to additional oxidation of the dimer bonds. An amorphous interfacial SiO_2 layer is likely to destroy these restrictions and thus lead to a lower injection barrier. □

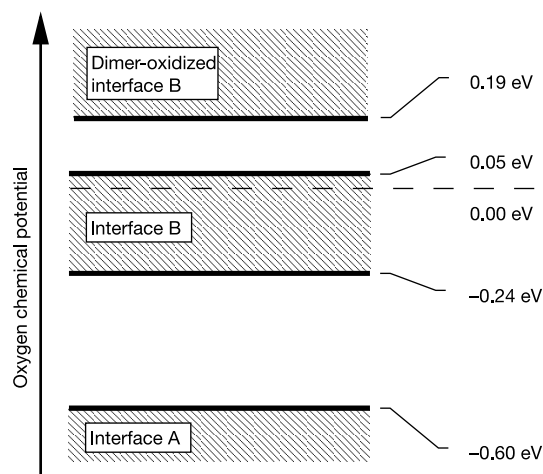


Figure 2 Phase diagram for interface oxidation. Shaded areas indicate the stability regions of the defect-free interfaces A and B and the dimer-oxidized variant of interface B. The blank regions separating them correspond to disordered structures with an oxygen content that increases with increasing chemical potential. The external parameter is the oxygen chemical potential. The zero of the chemical potential corresponds to the coexistence of bulk Si and SiO_2 (α -quartz) in thermal equilibrium.

Methods

We performed state-of-the-art electronic structure calculations and *ab initio* molecular dynamics simulations²⁵ based on density functional theory^{22,23,26} and the projector augmented wave method²⁷.

The calculations have been done on 5-layer slabs of silicon. The slab calculations included a vacuum region of at least $6 \text{ \AA}$ between repeated slabs. The relevant calculations of the interfaces are done in a (2×2) supercell. All structures are relaxed without symmetry constraints. The hydrogen-terminated silicon back plane has been kept frozen.

Our calculations used a plane wave cutoff of 30 Ry for the plane wave part of the wave function. We used the frozen-core approximation. Semi-core states of Sr and Ti have, however, been treated as valence electrons. We used the following sets of projector functions per angular momentum: $2s2p1d$ for oxygen, $2s2p1d$ for silicon, $3s2p2d$ for strontium and $2s2p2d$ for titanium.

For all calculations of Sr adsorption we used a grid with about 64 lateral k -points per

(1 × 1) surface unit cell. For the interfaces we used a grid corresponding to 16 *k*-points per (1 × 1) unit cell. For metallic structures we used the Mermin functional with a temperature of 1,000 K and the extrapolation to zero Kelvin proposed in ref. 28.

The band offsets have been derived by relating the plane wave part of the potential, averaged laterally, to band edges. The relative displacement between potential and band edges has been obtained from the epitaxially strained bulk materials.

The phase diagram in Fig. 2 was obtained from the total energies as a function of the oxygen chemical potential. A number of different, stoichiometric and non-stoichiometric, interface structures with varying oxygen content have been considered. The regions where one of the stoichiometric interfaces (A, B or the dimer-oxidized variant of B) is most stable are shaded. Tests for the upper boundary of interface A with larger, that is (4 × 4), unit cells confirmed that the phase boundaries obtained in (2 × 2) unit cell are reliable to about 0.03 eV.

Received 28 February; accepted 10 November 2003; doi:10.1038/nature02204.

- Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38**, 114–117 (1965).
- Semiconductor Industry Association. *International Technology Roadmap for Semiconductors* at (<http://public.itrs.net/>) (International SEMATECH, Austin, TX, 2001).
- Ishiwara, H. & Azuma, K. Oriented growth of SrTiO₃ films on Si(100) substrates using in-situ cleaning by excited hydrogen. *Mater. Res. Soc.* **116**, 369–374 (1988).
- Kado, Y. & Arita, Y. Heteroepitaxial growth of SrO films on Si substrates. *J. Appl. Phys.* **61**, 2398–2400 (1987).
- Mori, H. & Ishiwara, H. Epitaxial growth of SrTiO₃ films on Si(001) substrates using a focused electron beam evaporation method. *J. Appl. Phys.* **30**, 1415–1417 (1991).
- Kim, T. W. *et al.* Structural properties and interfacial layer formation mechanisms of PbTiO₃ thin films grown on p-Si substrates. *Appl. Phys. Lett.* **64**, 2676–2678 (1994).
- Alexe, M. Measurement of interface trap states in metal ferroelectric silicon heterostructures. *Appl. Phys. Lett.* **72**, 2283–2285 (1998).
- McKee, R. A., Walker, F. J. & Chisholm, M. F. Crystalline oxides on silicon: the first five monolayers. *Phys. Rev. Lett.* **81**, 3014–3017 (1998).
- McKee, R. A. & Walker, F. J. Process for depositing an oxide epitaxially onto a silicon substrate and structures prepared with the process. US patent 5,225,031 (United States Patent Office, 1993).
- McKee, R. A., Walker, F. J. & Chisholm, M. F. Physical structure and inversion charge at a semiconductor interface with a crystalline oxide. *Science* **293**, 468–471 (2001).
- Ashman, C., Först, C. J., Schwarz, K. & Blöchl, P. E. First-principles calculations of strontium adsorption on Si(001). *Phys. Rev. B* **69** (in the press).
- Fan, W. C., Wu, N. J. & Ignatiev, A. Observation of ordered structures of Sr on the Si(001) surface. *Phys. Rev. B* **42**, 1254–1257 (1990).
- Yao, X. *et al.* Initial stages of Ba adsorption on the Si(100)-(2 × 1) surface at room temperature. *Phys. Rev. B* **59**, 5115–5119 (1999).
- Wang, J. *et al.* Bonding and diffusion of Ba on a Si(001) reconstructed surface. *Phys. Rev. B* **60**, 4968–4971 (1999).
- Liang, Y., Gan, S. & Engelhard, M. First step towards the growth of single-crystal oxides on Si: formation of a two-dimensional crystalline silicate on Si(001). *Appl. Phys. Lett.* **79**, 3591–3593 (2001).
- Wang, J., Ooms, W. J. & Hallmark, J. A. Semiconductor structure having a crystalline alkaline earth metal oxide interface with silicon. US patent 6,238,459 (United States Patent Office, 2001).
- Droopad, R. *et al.* Epitaxial oxides on silicon grown by molecular beam epitaxy. *J. Cryst. Growth* **227–228**, 936–943 (2001).
- Robertson, J. & Peacock, P. W. Atomic structure, band offsets and hydrogen in high *k* oxide:silicon interfaces. *Mater. Res. Soc. Symp. Proc.* **747**, 99–111 (2002).
- Zhang, X., Demkov, A. A., Li, H., Hu, X. & Wei, Y. Atomic and electronic structure of the Si/SrTiO₃ interface. *Phys. Rev. B* **68**, 125323 (2003).
- Chambers, S. A. *et al.* Band offset and structure of SrTiO₃/Si(001) heterojunctions. *J. Vac. Sci. Technol. A* **19**, 934–939 (2001).
- Robertson, J. Band offsets of wide-band-gap oxides and implications for future electronic devices. *J. Vac. Sci. Technol. B* **18**, 1785–1791 (2000).
- Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev. B* **136**, 864–871 (1964).
- Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev. A* **140**, 1133–1138 (1965).
- Noland, T. A. Optical absorption of single-crystal strontium titanate. *Phys. Rev.* **94**, 724 (1954).
- Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471–2474 (1985).
- Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
- Gillan, M. G. Calculation of the vacancy formation energy in aluminum. *J. Phys. Condens. Matter* **1**, 689–711 (1989).

Acknowledgements We thank S. Chambers, M. Chisholm, W. Daum, A. Dimoulas, J. Fompeyrine, J.-P. Loquet, R. A. McKee, G. Norga and S. Stemmer for discussions. This work has been funded by the European Commission in the project 'Integration of Very High-*k* Dielectrics with CMOS Technology' (INVEST) and by the AURORA project of the Austrian Science Fund. Parts of the calculations have been performed on the computers of the 'Norddeutscher Verbund für Hoch- und Höchstleistungsrechnen' (HLRN).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to P.E.B. (peter.bloechl@tu-clausthal.de).

High-latitude controls of thermocline nutrients and low latitude biological productivity

J. L. Sarmiento¹, N. Gruber², M. A. Brzezinski³ & J. P. Dunne⁴

¹Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey 08544, USA

²IGPP and Department of Atmospheric Sciences, University of California at Los Angeles, Los Angeles, California 90095, USA

³Department of Ecology, Evolution and Marine Biology and the Marine Science Institute, University of California, Santa Barbara, California 93106, USA

⁴NOAA/Geophysical Fluid Dynamics Laboratory, PO Box 308, Forrestal Campus B Site, Princeton, New Jersey 08542, USA

The ocean's biological pump strips nutrients out of the surface waters and exports them into the thermocline and deep waters. If there were no return path of nutrients from deep waters, the biological pump would eventually deplete the surface waters and thermocline of nutrients; surface biological productivity would plummet. Here we make use of the combined distributions of silicic acid and nitrate to trace the main nutrient return path from deep waters by upwelling in the Southern Ocean¹ and subsequent entrainment into subantarctic mode water. We show that the subantarctic mode water, which spreads throughout the entire Southern Hemisphere^{2,3} and North Atlantic Ocean³, is the main source of nutrients for the thermocline. We also find that an additional return path exists in the northwest corner of the Pacific Ocean, where enhanced vertical mixing, perhaps driven by tides⁴, brings abyssal nutrients to the surface and supplies them to the thermocline of the North Pacific. Our analysis has important implications for our understanding of large-scale controls on the nature and magnitude of low-latitude biological productivity and its sensitivity to climate change.

The classical explanation for the observed nutrient distribution of the ocean in the low latitudes is that the downward flux of biogenic material from the surface of the ocean is balanced by upwelling of dissolved inorganic nutrients driven by vertical mixing in the main thermocline. This essentially one-dimensional view grew out of early theories of thermocline and thermohaline circulation that are no longer tenable. Estimates of the magnitude of vertical mixing in the main thermocline are about an order of magnitude smaller than required to explain the vertical profiles of tracers within this feature⁵. In addition, ocean model simulations of radiocarbon distribution show that balancing the formation of deep waters by upwelling through the main thermocline gives results that are inconsistent with observations¹. Instead, these studies suggest that a more likely return path for the deep water to the surface is in the Southern Ocean¹.

Subantarctic Mode Water (SAMW) has been identified as the main conduit of nutrients from the Southern Ocean to the upwelling regions of the equatorial Pacific and off South America⁶. The SAMW is a pycnostad (a layer of relatively uniform density) that originates in the thick wintertime mixed layers that ring the Southern Ocean⁷ (Fig. 1d). This belt, which is particularly strong eastward of the central Indian Ocean to the western South Atlantic, coincides with the Subantarctic Zone (SAZ) between the Subtropical Front at about 40–45°S and the Subantarctic Front at about 45–55°S, and appears to also include the Polar Front Zone (PFZ) between the Subantarctic Front and Polar Front just to the south (Fig. 1d). The SAMW pycnostad increases in density from $\sigma_\theta = 26.5$ (equivalent to 1,026.5 kg m⁻³) in the western Atlantic to $\sigma_\theta = 27.1$ in the southeast Pacific⁸ as it flows in an eastward circuit around the Southern Ocean.

One unusual characteristic of the waters found in this band is that they have high concentrations of nitrate (Fig. 1a), but low concentrations of silicic acid (Fig. 1b). We define a new tracer, $\text{Si}^* = [\text{Si}(\text{OH})_4] - [\text{NO}_3^-]$, whose concentrations of $-10 \mu\text{mol kg}^{-1}$ to $-15 \mu\text{mol kg}^{-1}$ in the SAMW formation regions (Fig. 1c) are the lowest we were able to find anywhere at the surface of the ocean. We use the low Si^* characteristic to trace the SAMW on the $\sigma_\theta = 26.8$ surface, which is the median of the range of densities that make up the SAMW (Si^* is nearly conserved in this water mass; see Methods). Figure 2a charts the reach of low- Si^* SAMW throughout the Southern Hemisphere subtropical gyres² and into the North Atlantic, where it is part of the upper water return flow of the global thermohaline circulation³.

In addition to being a powerful water mass tracer, Si^* is an indicator of nutrient status related to the requirements of diatoms. Diatoms with adequate light and nutrients (including iron) generally contain Si and N in a mole ratio of $\sim 1:1$ (ref. 9), which requires $\text{Si}^* \geq 0$. The presence of negative- Si^* SAMW waters at the base of the main thermocline (see depth in Fig. 2b) is associated with a silicic acid:nitrate ratio of ≤ 0.5 in the supply to surface waters (Fig. 2c; see Methods), which helps explain the prevalence of low diatom production over much of the world ocean⁹.

High nutrients in the surface waters of the Southern Ocean where SAMW forms cause this water type to be a significant source of nutrients to the rest of the ocean. To evaluate the influence of this supply on biological productivity in the rest of the ocean, we

performed a simulation with an ocean biogeochemistry general circulation model in which the Southern Ocean nutrient source was turned off by forcing nutrients towards zero at the surface of the ocean south of 30°S (see Methods). A comparison of this simulation to one in which surface nutrients were forced to observations everywhere suggests that the Southern Ocean nutrient supply accounts for about three-quarters of the biological production north of 30°S (Fig. 3).

The North Pacific Ocean is an important exception to the dominance of the SAMW influence. Figure 2a indicates that the main thermocline must be tapping directly into the high-nutrient waters of the deep ocean in this region¹⁰. The northernmost region of this basin is an area of upwelling; and wintertime cooling can make the water quite dense at the surface, particularly in the Sea of Okhotsk. The densest waters at the surface in the Sea of Okhotsk are only about $\sigma_\theta = 26.8$ (ref. 11), which is not high enough to allow deep waters to reach the surface. However, $\sigma_\theta = 26.8$ is the median density of upper North Pacific Intermediate Water¹² (NPIW), the formation of which is the result of a complex series of processes occurring in the Sea of Okhotsk^{4,13} (probably involving strong vertical mixing driven by tidal interactions in the Kurile Islands^{4,14}), as well as processes occurring in the 'mixed water region' between the Kuroshio and Oyashio currents^{12,15}. We suggest that the strong vertical mixing in the Kurile Islands region brings deep nutrients up into the thermocline at the depth of the NPIW as well as above it.

The influence of NPIW extends over the entire North Pacific to

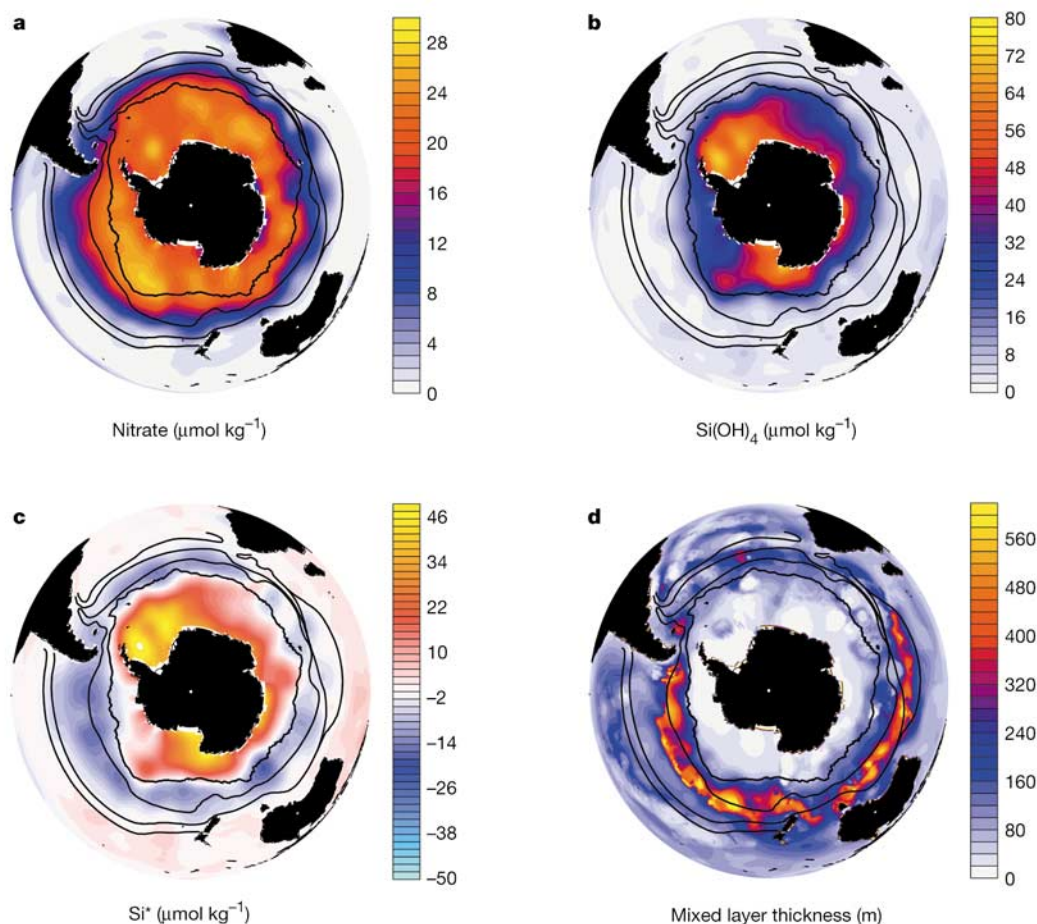


Figure 1 Polar stereographic maps of upper ocean nutrients and physics. **a**, Annual mean nitrate; **b**, annual mean silicic acid; **c**, annual mean Si^* ; and **d**, winter mixed layer thickness averaged over the July–September period. The southerly line denotes the mean position of the Polar Front²⁶. In sequence from south to north, the remaining lines

denote the position of the Subantarctic Front, the Southern Subtropical Front, and the Northern Subtropical Front²⁷. Nutrient data are from ref. 28; the mixed layer thickness data are from ref. 29.

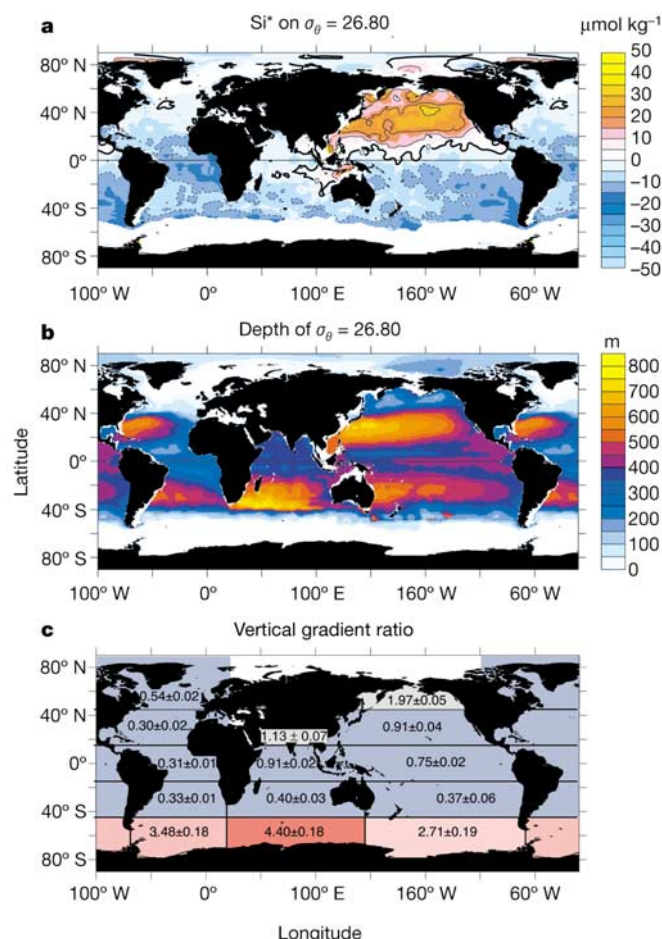


Figure 2 Global maps of nutrient properties mapped on the potential density surface $\sigma_{\theta} = 26.80$. The southern boundary is determined by where the annual mean density field outcrops at the surface of the ocean. **a**, Si^* . See Methods section for discussion of evidence that Si^* is conserved after the formation of SAMW in the SAZ. There are a few low-oxygen regions in the North Indian Ocean and eastern tropical Pacific where removal of nitrate as a consequence of denitrification decouples Si cycling from that of nitrate. We correct for this by using N^* (ref. 30), to define a corrected $\text{Si}^* = \text{Si}(\text{OH})_4 - \text{NO}_3^- - \delta(\text{N}^*)$ where δ is set to one if N^* is smaller than $-2 \mu\text{mol kg}^{-1}$ and to zero otherwise. Si^* is not conserved in the NPIW. **b**, Depth of the $\sigma_{\theta} = 26.80$ surface. The data for **a** and **b** are from ref. 28. **c**, The zonal mean of the ratio of the vertical gradient of silicic acid to that of nitrate between the surface 100 m and the next 100 m, calculated as described in the Methods section.

the Equator (Fig. 2a), where it accounts for 70% of the silicic acid supply for diatom production despite the fact that half the supply of nitrate comes from the Southern Hemisphere¹⁶. The influence of this water type can be seen at the surface of the ocean throughout the North Pacific, where the silicic acid to nitrate supply ratios are much higher than over the regions where SAMW influence dominates (Fig. 2c).

Model simulations in three-dimensional coupled biological–physical models consistently underpredict biological productivity in the North Pacific relative to observations¹⁷. This has been attributed to the failure to form NPIW with the right characteristics in these models¹⁷. Solving this problem may require developing a parameterization for the influence of tidally driven vertical mixing in the northwest Pacific.

There is also a modest amount of high- Si^* water in the eastern equatorial Indian Ocean and Bay of Bengal (Fig. 2a), where the supply ratio to the surface is also large (Fig. 2c). The former

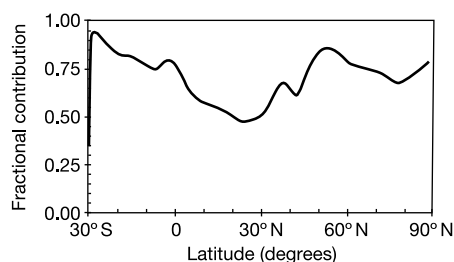


Figure 3 Predicted global zonal mean of the fractional contribution of Southern Ocean nutrient supply to global export production. Data obtained from an ocean biogeochemistry general circulation model (see model description in the Methods section). This figure is based on simulations carried out by R. Slater, with diagnostics provided courtesy of I. Marinov and A. Gnanadesikan.

probably results from North Pacific influence, but may also involve vertical mixing in the Indonesian Straits; the latter appears to be generated locally in the Bay of Bengal.

Given the importance of the SAMW and NPIW in the global thermocline nutrient cycle, we need to understand the processes that determine the properties of these water types. Previous work based on nitrogen isotopes suggests that surface nitrate in the SAZ and PFZ is carried northward by Ekman transport from the region to the south of the Polar Front¹⁸, where upwelling brings water with extremely high nutrient concentrations to the surface of the Southern Ocean (Fig. 1a and b). Organisms in the upwelling zone utilize the nutrients as these waters move towards the north (Fig. 4). Figure 1c shows that the waters furthest to the south have Si^* that goes as high as $50 \mu\text{mol kg}^{-1}$, but by the time they reach the Polar Front they have Si^* values of zero or less, with Si^* becoming increasingly negative towards the north within the Polar Front Zone and Subantarctic Zone (Figs 1c and 4). The preferential removal of silicic acid that would be required in order to generate such decreases in Si^* is a well known phenomenon (Fig. 2c), generally attributed to the influence of iron limitation in dramatically increasing the Si:N ratio of diatoms¹⁹. Studies in this region suggest that iron in sea ice and in upwelling deep water is sufficient to sustain intense diatom blooms but not high enough to prevent some manifestations of iron limitation, with the consequence that the Si:N ratio of diatoms is very high²⁰. Lateral mixing along isopycnals from the interior of the ocean plays a role in determining the surface nutrient distribution in the Southern Ocean as well²¹, but would not be able to provide the required upward supply of nutrients from the deep ocean to the SAMW formation region.

By contrast with the SAMW, the NPIW is rich in both nitrate and silicic acid. One might speculate that a difference in iron supply may play a role in mitigating the depletion of silicic acid relative to nitrate in this region. However, a more important contributor may be that NPIW formation, which occurs primarily by vertical mixing in the interior of the ocean, does not involve exposure at the surface of the ocean. The formation of NPIW thus avoids the biologically driven resetting of deep ocean nutrient ratios that occurs at the surface of the ocean during SAMW formation.

Our analysis has important implications for the impact of climate change on the global nutricline, biological productivity and the carbon cycle. The Southern Ocean has long been recognized as playing a central role in the global carbon cycle and biological productivity, and in the response of these to climate change^{22,23}. We show that it is specifically the processes determining the properties of SAMW, and the mechanisms leading to its formation in the SAZ, that are most crucial in determining how Southern Ocean processes affect the supply of nutrients to the main thermocline and low-latitude productivity. The SAZ is a region of deep mixed layers;

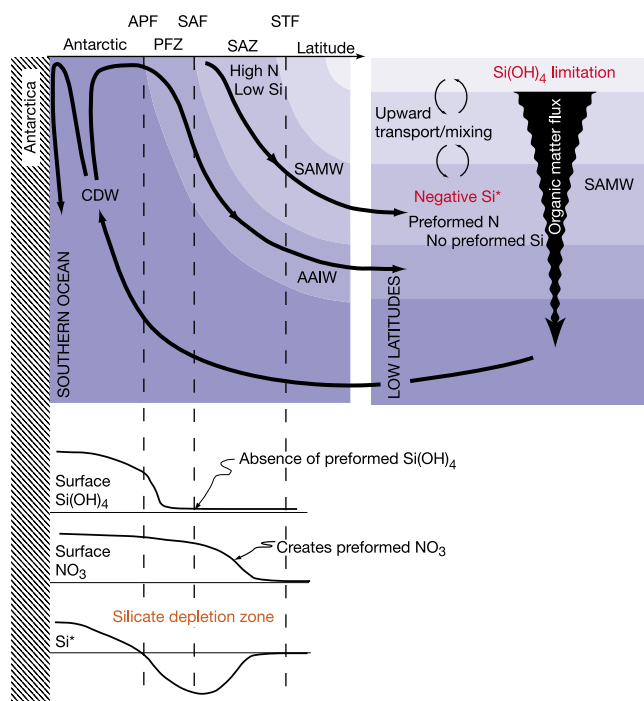


Figure 4 Southern Ocean control on thermocline nutrient concentrations. Conceptual diagram depicting the Southern Ocean physical and biological processes that form low- Si^* waters and feed them into the global thermocline. Top, water pathways; bottom, details of surface processes. Upper Circumpolar Deep Water (CDW) upwells to the surface in the Southern Ocean, and is transported to the north across the Antarctic Polar Front (APF) into the Polar Front Zone (PFZ), where Antarctic Intermediate Water (AAIW) forms, and then across the Subantarctic Front (SAF) into the Subantarctic Zone (SAZ), which is bounded to the north by the Subtropical Front (STF). Silicic acid is stripped out preferentially over nitrate as the water moves to the north, thus generating negative Si^* values. This negative- Si^* water is Subantarctic Mode Water (SAMW), which sinks into the base of the main thermocline and feeds biological production in the low latitudes.

models and observations suggest that these layers are formed by a combination of cooling of subtropical gyre water and northward wind driven Ekman transport²⁴. Both of these very probably changed during the ice ages and are likely to change with global warming, though the nature of the change is controversial. Our analysis also suggests that NPIW formation in the North Pacific is crucial in determining low-latitude productivity in the north and equatorial Pacific; NIPW formation thus merits greater attention than it has received in the past¹⁰. □

Methods

Conservation of Si^*

The usefulness of preformed Si^* as a tracer of the SAMW depends on the extent to which it is conserved. It will only be conserved if nitrate and silicic acid are recycled everywhere in a ratio of 1:1. Preferential remineralization of nitrate or of non-siliceous organisms, which is the behaviour that might be expected at the depth of the SAMW, will thus tend to reduce Si^* . Figure 2a does not show any significant reduction of Si^* away from the SAMW formation region, which provides qualitative support for conservation of this tracer.

We further examined this issue as follows. Both Si and N can be conceptually separated into preformed and remineralized components: $\text{NO}_3^- = \text{NO}_3^-_{\text{preformed}} + \Delta\text{NO}_3^-$, and $\text{Si(OH)}_4 = \text{Si(OH)}_4_{\text{preformed}} + \Delta\text{Si(OH)}_4$. Combining these gives $\text{Si}^* = \text{Si(OH)}_4 - \text{NO}_3^- = \text{Si(OH)}_4_{\text{preformed}} - \text{NO}_3^-_{\text{preformed}} + (\Delta\text{Si(OH)}_4 - \Delta\text{NO}_3^-)$. If remineralization were occurring with a 1:1 ratio, then $\Delta\text{Si(OH)}_4$ and ΔNO_3^- would be equal and $\text{Si}^* = \text{Si(OH)}_4_{\text{preformed}} - \text{NO}_3^-_{\text{preformed}}$. However, we expect some preferential remineralization of nitrate, so we further split ΔSi into a fraction that follows the 1:1 ratio and a term that contains the preferential remineralization, $\Delta\text{Si(OH)}_4 = \Delta\text{Si(OH)}_{4:1} + \Delta\text{Si(OH)}_{4:\text{preferential}}$. However, $\Delta\text{Si(OH)}_{4:1} = \Delta\text{NO}_3^-$, so we have $\text{Si}^* = \text{Si(OH)}_{4:\text{preferential}} - \text{NO}_3^-_{\text{preformed}} + \Delta\text{Si(OH)}_{4:\text{preferential}}$. On the $\sigma_\theta = 26.8$ surface, the preformed silicic acid concentration is close to 0, thus giving $\Delta\text{Si(OH)}_{4:\text{preferential}} = \text{Si}^* + \text{NO}_3^-_{\text{preformed}}$. We calculate preformed nitrate by subtracting the remineralized nitrate from the observed nitrate concentration,

$\text{NO}_3^-_{3\text{preformed}} = \text{NO}_3^-_{\text{observed}} - \Delta\text{NO}_3^-$, with remineralized nitrate calculated from the apparent oxygen utilization ($\text{AOU} = \text{O}_{2\text{saturation}} - \text{O}_{2\text{observed}}$) multiplied by the average N: O_2 stoichiometric ratio of 16:170, that is, $\Delta\text{NO}_3^- = \text{AOU} \cdot 16/170$. Except for the large North Pacific signal, which has a different preformed concentration, a plot of the preferential silicic acid distribution on the $\sigma_\theta = 26.8$ surface is within $\pm 5 \mu\text{mol kg}^{-1}$ of 0, demonstrating that our assumption of minimal impact of preferential remineralization on the Si^* distribution is a good one at the depth of the SAMW.

Export ratio

The opal to nitrogen export flux ratio is obtained from the water column nutrient data synthesis compiled from WOCE, SAVE, TTO and GEOSECS (see list of cruises in ref. 25); we used the method of ref. 25 to calculate the opal to organic carbon export flux ratio²⁵. Briefly, ref. 25 shows how the export flux ratio can be calculated from the vertical concentration gradient ratio of the precursor inorganic nutrients if certain assumptions are met. As in that study, we calculate the concentration gradient ratio by taking the mean over the top 100 m for the surface concentration, and the mean over 100 to 200 m for the thermocline concentration. The 'most likely' export flux ratio is then obtained from

$$\left(\frac{J_{\text{opal}}}{J_{\text{organic nitrogen}}} \right) = \frac{\sum (\text{Si(OH)}_4)_{100-200\text{m}} - (\text{Si(OH)}_4)_{0-100\text{m}}}{\sum ((\text{NO}_3^-)_{100-200\text{m}} - (\text{NO}_3^-)_{0-100\text{m}})}$$

As in ref. 25, we use the bootstrap method to estimate the nonparametric standard deviation of the mean. We construct 10,000 trial data sets from the original data set by random selection with replacement (that is, repeatedly randomly selecting from the same set of observations). The 'best' estimate of the gradient is the median of all the individual estimates, and the 95% confidence limits come directly from the 2.5% and 97.5% tails of the resulting distribution.

Contribution of Southern Ocean to export production

The model simulations of the influence of the Southern Ocean on ocean biological production are carried out in the KVLW + AILW ocean biogeochemical general circulation model discussed in ref. 17. Biological uptake of nutrients is obtained by forcing model predicted nutrients towards observed nutrients in the upper 75 m of the model. The model includes production and remineralization of both dissolved and particulate organic matter. We shut off the Southern Ocean nutrient supply to low latitudes by forcing surface nutrients towards zero south of 30°S such that the water sinking out of the surface has little or no nutrients in it. We note that the results we obtain are very sensitive to the model used, and that we forced nutrients towards zero everywhere south of 30°S , not just in the SAMW formation regions.

Received 11 May; accepted 8 October 2003; doi:10.1038/nature02127.

1. Toggweiler, J. R. & Samuels, B. in *The Global Carbon Cycle* (ed. Heimann, M.) 333–366 (Springer, Berlin, 1993).
2. McCartney, M. S. The subtropical recirculation of mode waters. *J. Mar. Res.* **40** (suppl.), 427–464 (1982).
3. Sloyan, B. M. & Rintoul, S. R. Circulation, renewal, and modification of Antarctic Mode and Intermediate Water. *J. Phys. Oceanogr.* **31**, 1005–1030 (2001).
4. Talley, L. D. An Okhotsk Sea water anomaly: Implications for ventilation in the North Pacific. *Deep-Sea Res.* **38** (suppl. 1), S171–S190 (1991).
5. Ledwell, J. R., Watson, A. J. & Law, C. S. Evidence for slow mixing across the pycnocline from an open-ocean tracer-release experiment. *Nature* **364**, 701–703 (1993).
6. Toggweiler, J. R., Dixon, K. & Broecker, W. S. The Peru upwelling and the ventilation of the South Pacific thermocline. *J. Geophys. Res.* **96**, 20467–20497 (1991).
7. McCartney, M. S. in *A Voyage of Discovery* (ed. Angel, M. V.) 103–119 (Supplement to Deep-Sea Research, George Deacon 70th Anniversary Volume, Pergamon, Oxford, 1977).
8. Hanawa, K. & Talley, L. D. in *Ocean Circulation and Climate* (eds Siedler, G. & Church, J.) 373–386 (Academic, San Diego, 2001).
9. Ragueneau, O. *et al.* A review of the Si cycle in the modern ocean: Recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy. *Glob. Planet. Change* **26**, 317–365 (2000).
10. Tsunogai, S. The Western North Pacific playing a key role in global biogeochemical cycles. *J. Oceanogr.* **58**, 245–257 (2002).
11. Reid, J. L. On the total geostrophic circulation of the Pacific Ocean: Flow patterns, tracers, and transport. *Prog. Oceanogr.* **29**, 263–352 (1997).
12. Talley, L. D. North Pacific Intermediate Water transports in the mixed water region. *J. Phys. Oceanogr.* **27**, 1795–1803 (1997).
13. Yasuda, I., Kouketsu, S., Katsumata, K. & Ohiwa, M. Influence of Okhotsk Sea Intermediate Water on the Oyashio and North Pacific Intermediate Water. *J. Geophys. Res.* **107**, doi:10.1029/2001JC001037 (2002).
14. Nakamura, T. *et al.* The generation of large-amplitude unsteady lee waves by subinertial K_1 tidal flow: A possible vertical mixing mechanism in the Kuril Straits. *J. Phys. Oceanogr.* **30**, 1601–1621 (2000).
15. Yasuda, I. *et al.* Hydrographic structure and transport of the Oyashio south of Hokkaido and the formation of North Pacific Intermediate Water. *J. Geophys. Res.* **106**, 6931–6942 (2001).
16. Dugdale, R. C. *et al.* Meridional asymmetry of source nutrients to the equatorial Pacific upwelling ecosystem and its potential impact on ocean-atmosphere CO_2 flux: a data and modeling approach. *Deep-Sea Res.* **49**, 2513–2531 (2002).
17. Gnanadesikan, A., Slater, R. D., Gruber, N. & Sarmiento, J. L. Oceanic vertical exchange and new production: A comparison between models and observations. *Deep-Sea Res.* **49**, 363–401 (2002).
18. Sigman, D. M., Altabet, M. A., McCorkle, D. C., Francois, R. & Fischer, G. The $\delta^{15}\text{N}$ of nitrate in the Southern Ocean: Consumption of nitrate in surface waters. *Glob. Biogeochem. Cycles* **13**, 1149–1166 (1999).
19. Franck, V. M., Brzezinski, M. A., Coale, K. H. & Nelson, D. M. Iron and silicic acid concentrations regulate Si uptake north and south of the Polar Frontal Zone in the Pacific Sector of the Southern Ocean. *Deep-Sea Res.* **47**, 3315–3338 (2000).

20. Brzezinski, M. A., Dickson, M.-L., Nelson, D. M. & Sambrotto, R. Ratios of Si, C and N uptake by microplankton in the Southern Ocean. *Deep-Sea Res. II* **50**, 619–633 (2003).
21. Pollard, R. T., Lucas, M. I. & Read, J. F. Physical controls on biogeochemical zonation in the Southern Ocean. *Deep-Sea Res. II* **49**, 3289–3305 (2002).
22. Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. & Manabe, S. Simulated response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393**, 245–249 (1998).
23. Sigman, D. M. & Boyle, E. A. Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* **407**, 859–869 (2000).
24. Speer, K., Rintoul, S. R. & Sloyan, B. The diabatic Deacon Cell. *J. Phys. Oceanogr.* **30**, 3212–3222 (2000).
25. Sarmiento, J. L. *et al.* A new estimate of the CaCO_3 to organic carbon export ratio. *Glob. Biogeochem. Cycles* **16**, doi:10.2929/2002GB00191 (2002).
26. Moore, J. K., Abbott, M. R. & Richman, J. G. Location and dynamics of the Antarctic Polar Front from satellite sea surface temperature data. *J. Geophys. Res.* **104**, 3059–3073 (1999).
27. Belkin, I. M. & Gordon, A. L. Southern Ocean fronts from the Greenwich meridian to Tasmania. *J. Geophys. Res.* **101**, 3675–3696 (1996).
28. Levitus, S. *et al.* *World Ocean Database 1998* Vol. 1, *Introduction* (NOAA NESDIS, Washington DC, 1998).
29. Kara, A. B., Rochford, P. A. & Hurlburt, H. E. Mixed layer depth variability over the global ocean. *J. Geophys. Res.* **108**, doi:10.1029/2000C000736 (2003).
30. Gruber, N. & Sarmiento, J. L. Global patterns of marine nitrogen fixation and denitrification. *Glob. Biogeochem. Cycles* **11**, 235–266 (1997).

Acknowledgements We thank I. Belkin and K. Moore for making their frontal paths in the Southern Ocean available to us. This paper benefited from comments by M. Bender, B. McNeil, D. Sigman, C. Sweeney and R. Toggeweiler. J.L.S. was supported by a NOAA Office of Global Programs grant to the Carbon Modeling Consortium for model development, and by an NSF grant for model and observational interpretations as part of the JGOFS Synthesis and Modeling Project. J.L.S. and J.D. were supported by a DOE Office of Science grant for the nutrient depletion scenarios. N.G. was supported by the DOE, and M.A.B. by the NSF.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.L.S. (jls@princeton.edu).

Stability of magnesite and its high-pressure form in the lowermost mantle

Maiko Isshiki^{1,2}, Tetsuo Irifune¹, Kei Hirose³, Shigeaki Ono⁴, Yasuo Ohishi², Tetsu Watanuki⁵, Eiji Nishibori⁶, Masaki Takata^{2,6} & Makoto Sakata⁶

¹Geodynamics Research Center, Ehime University, Matsuyama, Ehime 790-8577, Japan

²Spring-8/Japan Synchrotron Radiation Research Institute, 1-1-1 Kouto, Mikazuki, Sayo, Hyogo 679-5148, Japan

³Department of Earth and Planetary Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama, Tokyo 152-8551, Japan

⁴IFREE, Japan Marine Science & Technology Center, 2-15 Natsushima, Yokosuka, Kanagawa 237-0061, Japan

⁵Spring-8/Japan Atomic Energy Research Institute, 1-1-1 Kouto, Mikazuki, Sayo, Hyogo 679-5148, Japan

⁶Department of Applied Physics, Nagoya University, Furo-cho, Chikusa, Nagoya 464-8603, Japan

Carbonates are important constituents of marine sediments and play a fundamental role in the recycling of carbon into the Earth's deep interior via subduction of oceanic crust and sediments^{1–3}. Study of the stability of carbonates under high pressure and temperature is thus important for modelling the carbon budget in the entire Earth system. Such studies, however, have rarely been performed under appropriate lower-mantle conditions and no experimental data exist at pressures greater than 80 GPa (refs 3–6). Here we report an *in situ* X-ray diffraction study of the stability of magnesite (MgCO_3), which is the major component of subducted carbonates, at pressure and temperature conditions

approaching those of the core–mantle boundary. We found that magnesite transforms to an unknown form at pressures above ~115 GPa and temperatures of 2,100–2,200 K (depths of ~2,600 km) without any dissociation, suggesting that magnesite and its high-pressure form may be the major hosts for carbon throughout most parts of the Earth's lower mantle.

In equilibrium with silicates, the major carbonates in marine sediments (CaCO_3 calcite, CaCO_3 aragonite, and $\text{CaMg}(\text{CO}_3)_2$ dolomite) transform to dolomite and magnesite with increasing pressure to ~1–4 GPa (refs 7, 8), depending on temperature and bulk compositions. Dolomite has been shown to transform further, to magnesite-bearing assemblages in the presence of pyroxene and olivine at pressures of 20–50 GPa (ref. 3): accordingly, magnesite should be the only stable carbonate in the upper part of the mantle. Direct evidence of the presence of magnesite has also been demonstrated by its occurrence as inclusions in natural diamonds⁹, although it has been a matter of debate whether it is transported into the lower mantle via subduction^{3,10,11}.

The stability of MgCO_3 magnesite under lower-mantle conditions has been studied by using the diamond anvil cell (DAC); these studies showed that magnesite is stable at pressures to 80 GPa and temperatures of 2,000–2,500 K (refs 3–6). However, these studies were based on analyses of samples quenched to ambient conditions, or those at room temperature under high pressure, except for one study using Raman spectroscopy at 26 GPa and at 1,200 K (ref. 5), and no *in situ* X-ray diffraction studies under simultaneous high pressure and high temperature have been made. Moreover, the temperatures in most of these experiments were only poorly constrained and suffered large uncertainties (10–20% of the nominal values), and in some studies they were not even measured^{3,4}. The maximum pressures studied were also limited to 50–80 GPa, equivalent to depths of ~1,300–1,900 km in the mantle, and so the stability of magnesite throughout the mantle remains unresolved.

We made 12 runs of *in situ* X-ray measurements at high pressure and temperature, as summarized in Table 1. In each run, X-ray diffraction measurements were conducted at two or three different pressure–temperature conditions by increasing temperature at a fixed load. Magnesite was found to persist to pressures up to about 100 GPa, and at temperatures up to 3,000 K, and we did not see any evidence for the dissociation of MgCO_3 under these conditions (Fig. 1a–c). The unit-cell volume of magnesite was determined at high pressure after quenching to room temperature in each run. A least-squares fitting of

Table 1 **Experimental conditions and results**

Run number	P (GPa)	T (K)	t (min)	Results
Ms29	30.7	1,800	5	Magnesite
	30.2	2,300	5	Magnesite
Ms26	38.8	1,800	5	Magnesite
	41.8	2,200	5	Magnesite
Ms28	47.8	2,000	5	Magnesite
	50.9	2,500	5	Magnesite
Ms23	51.6	1,800	5	Magnesite
	54.8	2,200	5	Magnesite
	55.4	2,500	5	Magnesite
Ms7	59.1	1,800	5	Magnesite
	60.6	2,000	5	Magnesite
Ms22	62.5	1,800	5	Magnesite
Ms25	66.1	1,800	5	Magnesite
	63.0	2,400	5	Magnesite
Ms8	78.0	1,800	3	Magnesite
	81.6	2,500	3	Magnesite
	84.3	3,000	3	Magnesite
Ms16	96.9	2,000	10	Magnesite
Ms14	100.8	1,800	5	Magnesite
	100.1	2,200	5	Magnesite
Ms15	112.6	1,800	5	Magnesite
	114.4	2,200	5	Magnesite II (+magnesite)
Ms18*	116.3	2,100	10	Magnesite II (+magnesite)

* Al_2O_3 pressure medium was not used.

these data to the third-order Birch–Murnaghan equation of state yielded a zero-pressure bulk modulus of $K_0 = 103(2)$ GPa, assuming $K_0' = 5$, which is in reasonable agreement with that ($K_0 = 108(3)$ GPa) reported recently⁶.

We also observed X-ray diffraction peaks from magnesite at 112.6 GPa, 1,800 K, but noted that several new peaks appeared at the expense of those of magnesite when temperature was increased to 2,200 K, at 114.4 GPa, in the same run (Fig. 1d). An additional run was made at similar conditions (116.3 GPa and 2,100 K) without the Al_2O_3 pressure medium, and it was found that these peaks predominated over those of magnesite. These new peaks were present at room temperature after quenching under pressure (Fig. 1e), but disappeared when pressure was slightly decreased (Fig. 1f), and only the diffraction peaks of magnesite were observed in the recovered sample.

These new peaks did not correspond to any possible dissociation products of magnesite, and the presence of diffraction peaks for magnesite only in the recovered sample also precludes its dissociation at the high-pressure, high-temperature condition. Moreover, the formation of the same phase, regardless of the presence or absence of the corundum pressure medium, demonstrates that the surrounding Al_2O_3 is not involved in the reaction. Thus it is evident that magnesite transforms to a new high-pressure form (hereafter magnesite II) at pressures above ~ 115 GPa, at temperatures of 2,100–2,200 K.

Magnesite is expected to transform to the aragonite structure, as

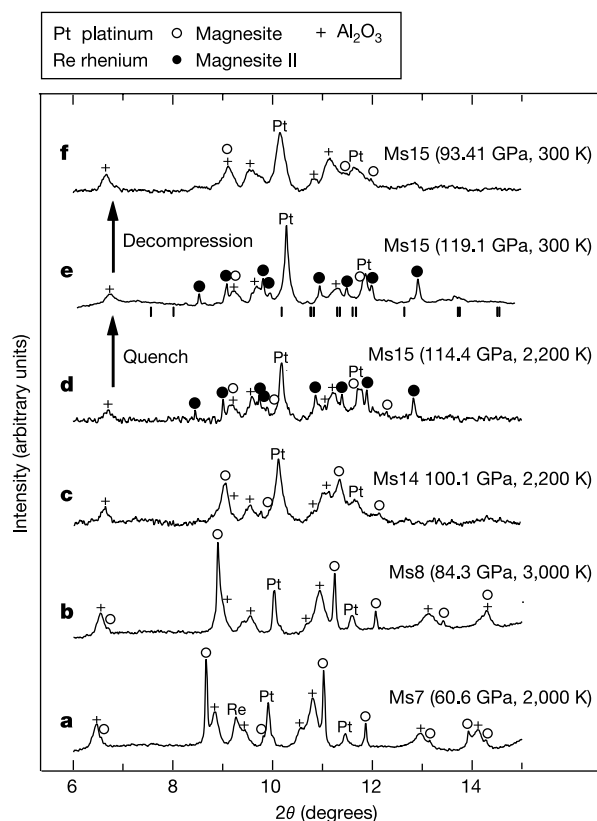


Figure 1 X-ray diffraction profiles with increasing pressure at the maximum temperatures in selected runs (a–d). Several new diffraction peaks corresponding to magnesite II were observed at pressures above ~ 115 GPa in Ms15 (d). These diffraction peaks survived when the run was quenched to the room temperature under pressure (e), but they were lost on release of pressure below 100 GPa (f). Additional peaks are mostly from platinum and Al_2O_3 . The vertical bars indicate the 2θ values expected for the major diffraction peaks ($I_0 > 20$) of MgCO_3 with the aragonite structure at 119.1 GPa and 300 K (see text).

is the case for CaCO_3 (calcite) at much lower pressures. We tried to index the observed diffraction peaks on the basis of aragonite structure; first we calculated the expected diffraction pattern for MgCO_3 with the aragonite structure at the ambient pressure by replacing Ca with Mg, taking the atomic scattering factors and M–O (where M is a metal ion) bond lengths into account. The diffraction pattern at 119.1 GPa was then calculated, assuming the axial compression of this phase to be close to that of magnesite⁶. However, we failed to assign these peaks to those based on the aragonite structure, as shown in Fig. 1e.

Some carbonates (SrCO_3 , PbCO_3 , BaCO_3) possessing the aragonite structure were reported to transform to unknown high-pressure phases with mutually closely related structures¹². However, it is unlikely that MgCO_3 adopts these structures at the pressure range of the present study, when the systematic relation between the ionic radii of divalent cations and the transition pressures of these phases¹³ is applied to the magnesium carbonate. CaCO_3 is estimated to adopt a corresponding structure at a pressure of ~ 130 GPa (ref. 13), so MgCO_3 would transform to a similar structure at still higher pressures, because of the much smaller ionic radius of Mg (than that of Ca) according to the above systematic relation.

The X-ray diffraction peaks of magnesite II can be reasonably indexed on the basis of its orthorhombic symmetry, although we were unable to refine the crystal structure of this phase, partly because of possible overlap of the diffraction peaks with those of corundum, platinum, magnesite and rhenium. The observed and calculated lattice spacing (d) values for magnesite II at 119.1 GPa after heating are shown in Table 2. We tentatively determined the unit-cell parameters as $a = 7.180(3)$ Å, $b = 5.027(10)$ Å, $c = 4.476(11)$ Å, and $V = 161.6(3)$ Å³, on the basis of orthorhombic symmetry and using eight diffraction peaks. If we assume the number of molecules in a unit cell to be $Z = 6$, we obtain a room-temperature density of 5.20 g cm^{−3} for the new phase at 119.1 GPa, which is about 17% denser than that of magnesite (4.46 g cm^{−3}) under the same pressure–temperature conditions. It should be noted, however, that alternative indexing of the diffraction peaks may be possible, and further study is needed to unambiguously identify the crystal structure of magnesite II.

Figure 2 shows the present experimental conditions and the results, and depicts a typical geotherm¹⁴ in the mantle and a predicted boundary for the dissociation of MgCO_3 (ref. 6). It is unlikely that the dissociation of magnesite into MgO and CO_2 occurs throughout the most parts of the lower mantle. It has been proposed that magnesite might decompose into an assemblage of MgO (periclase) + C (diamond) + O_2 under lower-mantle conditions, at pressures greater than 40 GPa (ref. 15), resulting in the formation of diamonds containing inclusions of magnesiowüstite¹⁶. However, the present experimental results demonstrate that such a reaction is also unlikely to occur along the geotherm, at least at pressures up to ~ 115 GPa.

There are several lines of evidence that suggest some crustal carbonates subduct to depths of at least 100 km in the mantle^{1,2}. Subducted slabs are significantly colder than the surrounding

Table 2 Observed and calculated d -values for magnesite II

hkl	d_{obs} (Å)	d_{cal} (Å)	$d_{\text{obs}} - d_{\text{cal}}$ (Å)
020	2.5189	2.5139	0.0050
120	2.3677	2.3723	−0.0046
021	2.1919	2.1920	−0.0001
310	2.1611	2.1612	−0.0001
112	1.9646	1.9667	−0.0022
221	1.8728	1.8708	0.0022
400	1.7953	1.7950	0.0003
401	1.6660	1.6661	−0.0001

Indexed by an orthorhombic symmetry with $a = 7.180(3)$ Å, $b = 5.027(10)$ Å, $c = 4.476(11)$ Å, $V = 161.6(3)$ Å³, using radiation with $\lambda = 0.3751$ Å. Standard deviations in parentheses.

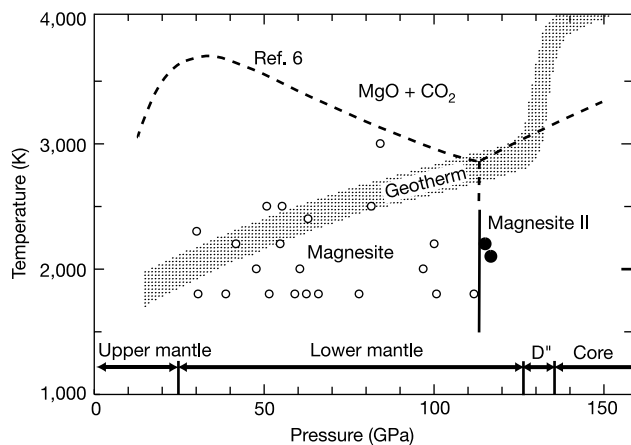


Figure 2 Possible phase relations of MgCO_3 in the deep mantle. The dissociation boundary is based on a calculation assuming $dK/dT = -0.013 \text{ GPa K}^{-1}$ (where K is the bulk modulus) for magnesite, which provided the minimum dissociation temperatures in the calculations of ref. 6. A typical geotherm based on ref. 14 is illustrated for comparison.

mantle, and accordingly at least some fractions of the carbonates may survive and be transported further down to the deep mantle without dissociation into a CO_2 -bearing assemblage. Images from seismic tomography suggest that some subducted slabs stagnate around the 660-km discontinuity¹⁷, whereas others seem to penetrate deeply into the lower mantle¹⁸. Moreover, parts of the stagnant slabs may ultimately sink into the lower mantle after they have accumulated to a critical size to form a megalith structure¹⁹. Thus magnesite is likely to be transported further into the lower mantle, there to transform into magnesite II at $\sim 115 \text{ GPa}$, equivalent to a depth of $\sim 2,600 \text{ km}$.

The present experiments suggest that magnesite II is the major host for oxidized carbon in the D'' layer, where its density is comparable to that estimated on seismological grounds²⁰ ($5.3\text{--}5.5 \text{ g cm}^{-3}$). It is acknowledged that the abundance of carbon near the surface of the Earth is significantly lower than that expected from a cosmochemical ground²¹, and some of such missing carbon⁴ may be stored in this phase at the base of the mantle. It is also possible that a certain amount of carbon is incorporated into the iron core, following the reaction MgCO_3 (magnesite II) + $5\text{Fe} \leftrightarrow \text{Fe}_3\text{C} + 2\text{FeO} + \text{MgO}$ (ref. 22), or forms a solid solution with (Mg,Fe)O magnesioiwüstite³ in the D'' layer, representing alternative hosts for the Earth's missing carbon. On the other hand, Fig. 2 shows that magnesite II may dissociate to $\text{MgO} + \text{CO}_2$ in the D'' layer, as a result of the dramatic temperature increase with depth in this region, associated with formation of a thermal boundary layer between the silicate mantle and the iron-dominant core. Release of CO_2 from magnesite II may trigger partial melting of materials present in the D'' layer, which should contribute to initiate rising plumes from this region²³. CO_2 formed in this way would be entrained by the hot plumes, and react with magnesioiwüstite in the plumes at some depths in the lower mantle to form carbonates according to Fig. 2. Such carbonates would again dissociate into CO_2 -bearing assemblages upon rising into the uppermost part of the mantle, causing kimberlite and other carbonate-rich volcanism. □

Methods

Starting materials

We used natural magnesite with a composition of $(\text{Mg}_{0.994}\text{Ca}_{0.006})\text{CO}_3$ as the starting material for the high-pressure runs. The sample was ground into a fine powder in a ceramic mortar and mixed with a small amount ($<10 \text{ wt\%}$) of platinum powder,

which was used as a laser-radiation absorber as well as an internal pressure marker. The pressure medium was an Al_2O_3 powder, which also worked as a thermal insulator between the sample and the diamond anvil. The sample was sandwiched by the Al_2O_3 powders, and enclosed in a small ($60\text{--}100 \mu\text{m}$ in diameter) hole within a Re gasket.

Experimental procedure under high pressure and high temperature

High-pressure and high-temperature experiments were carried out using the double-sided laser-heated DAC technique²⁴. The sample was first compressed to the target pressure, and then heated with a multimode continuous Nd:YAG laser. The laser beam was focused to $70\text{--}80 \mu\text{m}$ in diameter, and the sample was irradiated with the beams via each of a pair of diamond anvils. *In situ* X-ray diffraction measurements were carried out at beamline 10XU of SPring-8 at pressures of $30\text{--}120 \text{ GPa}$ and temperatures of $1,800\text{--}3,000 \text{ K}$. The incident X-ray beam was monochromatized to wavelengths of 0.3571 or $0.4130 \text{ \AA}$ and collimated to $20 \mu\text{m}$ in diameter, and the diffracted beam was detected with an imaging plate. Typical exposure time was $3\text{--}10 \text{ min}$ for each pressure-temperature condition, and the phases present were identified from the diffraction patterns. The diffraction images were integrated with the standard software, and the obtained one-dimensional diffraction patterns were treated with profile-fitting software using the LeBail method to identify the phases present and to determine the lattice parameters.

Precision of pressure and temperature determination

The sample temperature was measured on one side by using Wien's black body radiation law on the basis of thermal radiation from the sample. The temperature uncertainty due to fluctuations during the heating is about $\pm 100 \text{ K}$. The pressure at high temperature was determined from the observed unit-cell volumes of Pt using an equation of state²⁵, while the pressure upon compression at the room temperature was estimated based on the ruby fluorescence method²⁶. The precision of pressure determination, owing to temperature fluctuations at high temperature, was $\pm 0.5\text{--}0.7 \text{ GPa}$, but may suffer uncertainties of the order of $\sim 2 \text{ GPa}$ because of the presence of significant pressure and temperature gradients in the sample.

Received 23 April; accepted 4 November 2003; doi:10.1038/nature02181.

- Chopin, C. Coesite and pure pyrope in high-grade blueschists of the western Alps: a first record and some consequences. *Contrib. Mineral. Petrol.* **86**, 107–118 (1984).
- Becker, H. & Altherr, R. Evidence from ultra-high-pressure marbles for recycling of sediments into the mantle. *Nature* **358**, 745–748 (1992).
- Biellmann, C., Gillet, P., Guyot, F., Peyronneau, J. & Reynard, B. Experimental evidence for carbonate stability in Earth's lower mantle. *Earth Planet. Sci. Lett.* **118**, 31–41 (1993).
- Katsura, T. *et al.* Stability of magnesite under the lower mantle conditions. *Proc. Jpn Acad. B* **67**, 57–60 (1991).
- Gillet, P. Stability of magnesite (MgCO_3) at mantle pressure and temperature conditions: A Raman spectroscopic study. *Am. Mineral.* **78**, 1328–1331 (1993).
- Fiquet, G. *et al.* Structural refinements of magnesite at very high pressure. *Am. Mineral.* **87**, 1261–1265 (2002).
- Egger, D. H., Kushiro, I. & Holloway, J. R. Stability of carbonate minerals in a hydrous mantle. *Carnegie Inst. Wash. Yr Bk* **75**, 631–636 (1976).
- Wyllie, P. J. & Huang, W. L. Carbonation and melting reactions in the system $\text{CaO-MgO-SiO}_2\text{-CO}_2$ at mantle pressure with geophysical and petrological applications. *Contrib. Mineral. Petrol.* **54**, 79–107 (1976).
- Wang, A., Pasteris, J. D., Meyer, H. O. A. & Dele-Duboi, M. L. Magnesite-bearing inclusion assemblage in natural diamond. *Earth Planet. Sci. Lett.* **141**, 293–306 (1996).
- Domanik, K. J. & Holloway, J. R. Experimental synthesis and phase relations of phengitic muscovite from 6.5 to 11 GPa in a calcareous metapelite from the Dabie Mountains, China. *Lithos* **52**, 51–77 (2000).
- Hamouda, T. High-pressure melting of carbonated eclogite and experimental constraints on carbon recycling and storage in the mantle. *Earth Planet. Sci. Lett.* **214**, 357–368 (2003).
- Lin, C.-C. & Liu, L.-G. High pressure phase transformations in aragonite-type carbonates. *Phys. Chem. Miner.* **24**, 149–157 (1997).
- Lin, C.-C. & Liu, L.-G. Post-aragonite phase transitions in strontianite and cerussite—a high-pressure Raman spectroscopic study. *J. Phys. Chem. Solids* **58**, 977–987 (1997).
- Brown, J. M. & Shankland, T. J. Thermodynamic parameters in the earth as determined from seismic profiles. *Geophys. J. R. Astron. Soc.* **66**, 579–596 (1981).
- Liu, L.-G. Genesis of diamonds in the lower mantle. *Contrib. Mineral. Petrol.* **134**, 170–173 (1999).
- Harte, B. & Harris, J. W. Lower mantle mineral associations preserved in diamonds. *Mineral. Mag.* **58A**, 384–385 (1994).
- Fukao, Y., Obayashi, M., Inoue, H. & Nishii, M. Subducting slabs stagnant in the mantle transition zone. *J. Geophys. Res.* **97**, 4809–4822 (1991).
- Van der Hilst, R. D., Engdahl, E. R., Spakman, W. & Nolet, G. Tomographic imaging of subducted lithosphere below northwest Pacific island arcs. *Nature* **353**, 37–42 (1991).
- Ringwood, A. E. & Irifune, T. Nature of the 650-km seismic discontinuity: implications for mantle dynamics and differentiation. *Nature* **331**, 131–136 (1988).
- Dziewonski, A. M. & Anderson, D. L. Preliminary reference earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
- Anders, E. & Owen, T. Mars and Earth: origin and abundance of volatiles. *Science* **198**, 453–465 (1977).
- Scott, H. P., Williams, Q. & Knittle, E. Stability and equation of state of Fe_3C to 73 GPa: Implications for carbon in the Earth's core. *Geophys. Res. Lett.* **28**, 1875–1878 (2001).
- Zhao, D. Seismic structure and origin of hotspots and mantle plumes. *Earth Planet. Sci. Lett.* **192**, 251–265 (2001).
- Watanuki, T., Shimomura, O., Yagi, T., Kondo, T. & Isshiki, M. Construction of laser-heated diamond anvil cell system for *in situ* x-ray diffraction study at SPring-8. *Rev. Sci. Instrum.* **72**, 1289–1292 (2001).

25. Jamieson, J. C., Fritz, J. N. & Manghnani, M. H. in *High-Pressure Research in Geophysics* (eds Akimoto, S. & Manghnani, M. H.) 27 (Center for Academic Publications Japan, Tokyo, 1982).
26. Mao, H. K., Bell, P. M., Shaner, J. W. & Steinberg, D. J. Specific volume measurements of Cu, Mo, Pd, and Ag and calibration of the ruby R₁ fluorescence pressure gauge from 0.06 to 1 Mbar. *J. Appl. Phys.* **49**, 3276–3283 (1978).

Acknowledgements We thank A. Kurio, M. Murakami, Y. Kuwayama and N. Sata for discussions and technical assistance. We also thank T. Katsura for providing the magnesite starting material. M.I. thanks T. Yagi for advice on DAC techniques. The *in situ* X-ray observations were conducted at BL10XU, SPring-8.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.I. (issihiki@sci.ehime-u.ac.jp) or T.I. (irifune@dpc.ehime-u.ac.jp).

Unmatched tempo of evolution in Southern African semi-desert ice plants

C. Klak¹, G. Reeves² & T. Hedderson¹

¹Bolus Herbarium, Department of Botany, University of Cape Town, 7701 Rondebosch, South Africa

²Leslie Hill Molecular Systematics Laboratory, Kirstenbosch Research Centre, National Botanical Institute Cape Town, Private Bag X7, 7735 Claremont, South Africa

The Succulent Karoo is an arid region, situated along the west coast of southern Africa. Floristically this region is part of the Greater Cape Flora¹ and is considered one of the Earth's 25 biodiversity hotspots². Of about 5,000 species occurring in this region, more than 40% are endemic³. Aizoaceae (ice plants) dominate the Succulent Karoo both in terms of species numbers (1,750 species in 127 genera) and density of coverage^{3,4}. Here we show that a well-supported clade within the Aizoaceae, representing 1,563 species almost exclusively endemic to southern Africa, has diversified very recently and very rapidly. The estimated age for this radiation lies between 3.8 and 8.7 million years (Myr) ago, yielding a per-lineage diversification rate of 0.77–1.75 per million years. Both the number of species involved and the tempo of evolution far surpass those of any previously postulated continental or island plant radiation^{5–7}. Diversification of the group is closely associated with the origin of several morphological features and one anatomical feature. Because species-poor clades lacking these features occur over a very similar distribution area, we propose that these characteristics are key innovations that facilitated this radiation.

Most species of Aizoaceae exhibit leaf succulence, whereas stem succulents are comparatively rare. The range of life forms, particularly in the highly succulent Aizoaceae, is remarkable. It includes annuals, perennials, trailing woody plants, shrubs and even small trees, as well as highly compact succulent forms (sometimes reduced to just two leaves), geophytes and a few stem succulents with deciduous leaves. The diversity of both life form and number of species has been largely attributed to a complex interaction between the availability of numerous and diverse niches⁸ and strong genetic drift caused by very low rates of gene exchange between populations⁴. Because most species are found within the arid winter rainfall area, the onset of the winter rainfall regime (about 5 Myr ago) is also thought to have heralded the beginning of diversification in the Aizoaceae. In the past, the existing, although very scanty,

fossil evidence has been interpreted as providing support for the theory that between 10 and 5 Myr ago there was a major change in the composition of the flora in the region, from a moister to a much drier climate^{9,10}. However, arid to semi-arid conditions interspersed with wetter periods dominated the climate of areas adjacent to the Succulent Karoo, such as the Namib, since the Cretaceous period about 80 Myr ago^{11,12}. It is therefore possible that the Aizoaceae were well represented long before the onset of the winter rainfall regime.

In the absence of a good fossil record, molecular phylogenies are now used increasingly^{5,6} to evaluate competing hypotheses about the timing of diversification. A key feature of all the DNA regions we investigated is the low level of divergence between representatives of the clade incorporating most members of the subfamily Ruschioideae. This clade represents a large group of about 1,563 species in 101 genera and thus constitutes about 85% of the Aizoaceae (Fig. 1). The small number of substitutions coupled with the high species diversity suggests very recent and very rapid diversification. The number of species involved in this radiation far exceeds that of any other group in which recent and rapid radiation has been postulated^{5–7}. Contrary to an earlier hypothesis suggesting that all highly succulent Aizoaceae (subfamilies Mesembryanthemoideae and Ruschioideae) are of recent origin⁴, we show here that only the core Ruschioideae has radiated recently, with an estimated age of between 3.8 and 8.7 Myr. Our estimates of the onset of the radiation therefore fall within the proposed time frame of increased aridification and opening of the winter-rainfall meganiche about 5 Myr ago^{9,10}.

The estimated per lineage diversification rate per million years lies between 0.77–1.75 (assuming zero extinction) and 0.58–1.32 (when incorporating a high rate of extinction¹³). Thus, the tempo of radiation in the core Ruschioideae is considerably higher than that for any other continental (angiosperm families, median of 0.12 and maximum of 0.39 species per Myr¹⁴; angiosperm orders, maximum of 0.76 species per Myr (ref. 13)) or island (Hawaiian silversword alliance, 0.56 ± 0.17 species per Myr (ref. 7)) plant radiation. This estimate also compares favourably with animal diversification rates (hexapods, 0.05 species per Myr (ref. 15); Neogene horses, 0.5–1.4 species per Myr (ref. 16); Hawaiian island drosophilids, 1.21 species per Myr; Lake Tanganyika cichlids, 0.75–1.49 species per Myr (ref. 17)). Even higher diversification rates have been reported for some cichlid radiations¹⁷. However, our study assumes a constant rate of diversification over a time interval of 3.8–8.7 Myr, and it is likely that the core Ruschioideae have undergone intervals of even higher diversification rates¹⁸. The question of features or conditions that might have facilitated this radiation is important. The two lineages basal to the core Ruschioideae, as well as the next closely related subfamily, the Mesembryanthemoideae, all contain far fewer species: the Mesembryanthemoideae consists of about 100 species, and the two lineages basal to the core Ruschioideae each consist of 11 species. This indicates a large imbalance in net speciation rates between these clades and the core Ruschioideae. Yet the species-poor lineages also consist exclusively of highly succulent plants. Furthermore, they share with the core Ruschioideae both a similar distribution area and a similar dispersal strategy. Therefore, climatic or ecological factors alone cannot explain the remarkable speciation burst. We propose that several morphological characters constitute key innovations that have facilitated the major radiation of the core Ruschioideae. One important invention seems to be the recently reported wide-band tracheids, which are found (with few exceptions) in the core Ruschioideae only and are absent from all taxa that are sister to this clade. These specialized cells are thought to prevent the collapse of the primary wall and have therefore been associated with adaptations to withstand water stress^{19,20}. Taxa that possess these cells are likely to have an increased chance of survival in an arid environment over those taxa that lack them. A further major change in morphology associated with the node defining the

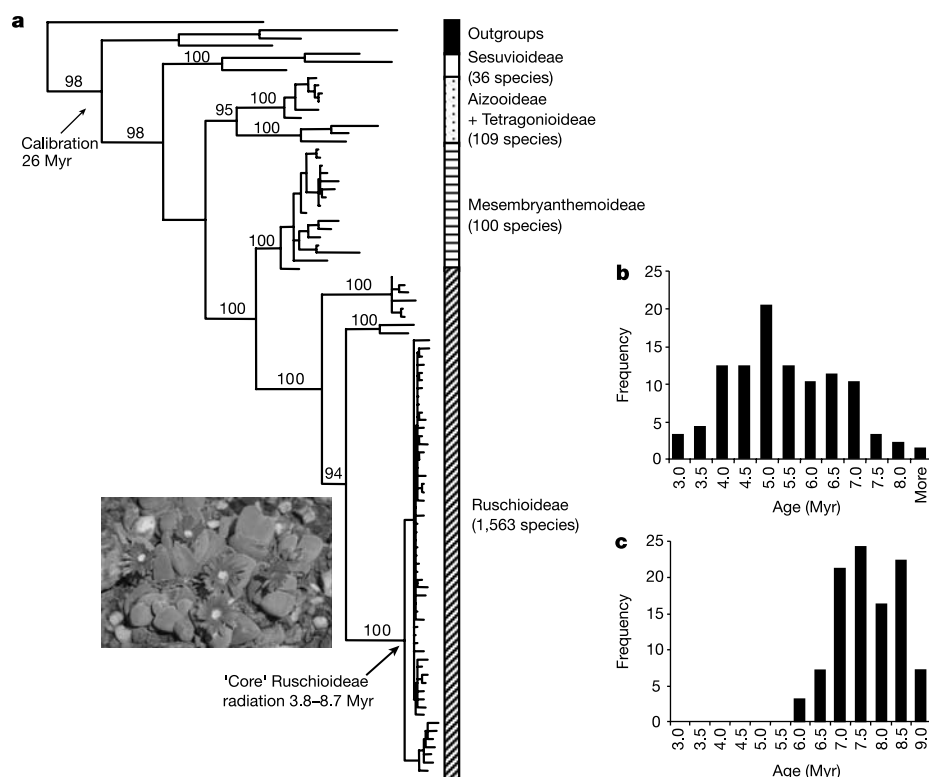


Figure 1 Parsimony analysis. **a**, One of the 5,000 equally parsimonious trees selected randomly from the parsimony analysis of the plastid data (tree length 1,413 steps) with bootstrap percentages indicated above the branches. Branch lengths are proportional to the changes in the plastid data. Calibration of the split between Nyctaginaceae and Aizoaceae + Phytolaccaceae at 26 Myr gave an estimated age for the 'core'

Ruschioideae clade of 3.8–8.7 Myr. The photograph shows a typical member of the 'core' Ruschioideae, *Gibbaeum haaglenii* H. E. K. Hartmann. **b**, ITS bootstrap distribution of age estimates. Mode = 3.8; s.d. = 3.2 Myr. **c**, Plastid bootstrap distribution of age estimates. Mode = 8.7; s.d. = 0.7 Myr.

species-rich clade is leaf shape. Flat or flattened leaves are common in all of the basal clades, whereas the leaves typically show a cylindrical or trigonous shape in the species-rich clade. This reduction in leaf surface area provides a further advantage in reducing water loss under dry conditions. Further xeromorphic modifications (particularly those of the leaves) are characteristic within the core Ruschioideae, but these are not directly associated with the onset of the radiation that we have dated in this study.

In addition, hygrochastic capsules, which open only when moistened to release a portion of their seeds in rain, might have had a function in the radiation of the core Ruschioideae. Although hygrochastic capsules are found in both the species-poor (such as the Mesembryanthemoideae) and the species-rich clades (core Ruschioideae), the morphology of the capsules found in the species-rich clade is far more specialized. Experimental removal of some or all of the specialized structures in Ruschioid capsules results in the retention of fewer seeds²¹, suggesting that these structures spread seed dispersal through time, thus increasing the probability of seed release during rainfall events of adequate size and duration to ensure seed germination and the subsequent survival of seedlings. Members of the Mesembryanthemoideae always lack these structures and their seeds are consequently lost in a few rainfall events. In contrast, fruits found in members of the core Ruschioideae release only a few seeds at a time. This is considered to confer advantages because the presence of seeds for subsequent rains is an effective dispersal strategy in regions with mostly erratic rainfall. The coupling of both dispersal and germination (although applicable to both the species-poor and species-rich clades) in turn reduces the possibility of secondary dispersal of seed, thus reducing gene-flow

distances and potentially promoting isolation of populations in space²¹. □

Methods

DNA sequencing

We used *rps16* intron and *trnL-F* sequences for 91 Aizoaceae species, representing all major lineages in the family, and four outgroup taxa (from the Molluginaceae, Phytolaccaceae and Nyctaginaceae)²². Internal transcribed spacer (ITS) sequences were also produced for 34 ingroup and 3 outgroup taxa.

Phylogenetic analyses

We analysed two matrices (the first consisting of combined *rps16* intron and *trnL-F* sequences and the second of ITS sequences) using the parsimony algorithm of PAUP* (ref. 23). Each data set was subjected to 1,000 random addition replicates using equal weights²⁴ with TBR branch swapping. Internal support was assessed using 1,000 bootstrap replicates²⁵. The topologies retrieved in each analysis were similar with respect to the recovery of, and relationships between, the major lineages. Minor exceptions concerned the placement of *Tetragonia* and *Dorotheanthus* + *Claretum* in the ITS analysis. Because we considered the more complete analysis including 95 taxa for two regions to be our best estimate of phylogenetic relationships, we used this as our best estimate of the phylogeny. High bootstrap percentages were recovered for all major lineages for both analyses, and differences in equally parsimonious trees for both data sets were restricted to terminal groupings mainly within the core Ruschioideae.

Age and diversification of core Ruschioideae

Maximum likelihood (ML) branch lengths were fitted using a three-parameter model; the transition–transversion ratio, base frequencies and gamma distribution of rate variability between sites were all estimated from the data. We chose this model over less complex alternatives with a likelihood ratio (LR) test in which the statistic was twice the difference in log likelihood scores between the simpler model and the more complex model. At each step we chose the more complex model if it provided a significantly better fit to the data. The hypothesis of rate constancy was rejected for both data sets using an LR test between a rate-constrained tree (forcing a molecular clock in PAUP*) and a rate-unconstrained tree. Ultrametric trees were therefore produced from both parsimony and ML branch lengths

with Sanderson's²⁶ method of non-parametric rate smoothing (NPRS) implemented in TreeEdit version 1.0 alpha 4-61 (ref. 27). In the absence of reliable fossil data^{28,29}, all trees were calibrated using an estimated age of 26 Myr for the split between Aizoaceae + Phytolaccaceae and its nearest relative, the Nyctaginaceae. We took this date from a DNA-sequence-based angiosperm phylogeny using ML branch lengths³⁰. In addition, to account for the error in the calibration, we recalculated the age of the core Ruschioideae and diversification rate using the most conservative date estimated for the split between Aizoaceae + Phytolaccaceae and Nyctaginaceae (that is, 30 Myr using parsimony branch lengths and DELTRAN optimization)³⁰. The standard error in age estimates for each data set with ML branch lengths was also estimated using 100 bootstrap matrices and one of the most equally parsimonious trees found from the initial heuristic searches (*Tetragonia* and *Dorotheanthus* + *Cleretum* once again constrained for the ITS analysis). With a calibration age of 26 Myr, the estimated age of the core Ruschioideae radiation from the plastid data set was 6.2 Myr with ML branch length (mode of bootstrap distribution 8.7 ± 0.7 Myr; Fig. 1) and 6.4 Myr with parsimony branch lengths; and from the ITS data set 3.1 Myr with ML branch lengths (mode of bootstrap distribution 3.8 ± 3.2 Myr; Fig. 1) and 6.6 Myr with parsimony branch lengths. With a calibration age of 30 Myr, the mode of the bootstrap distribution was slightly higher (plastid data 9.5 ± 0.8 Myr and ITS data 4.4 ± 3.7 Myr with ML branch lengths).

We estimated the per-lineage rate of diversification per million years for the core Ruschioideae radiation from both plastid and ITS data sets as $(\ln N - \ln N_0)/T$ (ref. 7), where initial diversity $N_0 = 2$, N is existing diversity and T is estimated clade age (we used the mode of the bootstrap distribution of age estimates as T). Thus, with a calibration of 26 Myr the estimated per-lineage diversification rate per million years based on the plastid data was 0.77 and for ITS 1.75 (0.70 and 1.5 respectively for the plastid and ITS data with a calibration point of 30 Myr). Taking extinction into account with a high value of $\epsilon = 0.9$ (ref. 13), we recalculated diversification rates as 0.58 lineages per million years for the plastid data and 1.32 lineages per million years for the ITS data.

Received 5 August; accepted 11 November 2003; doi:10.1038/nature02243.

- Jürgens, N. A new approach to the Namib region. Part I: phytogeographic subdivision. *Vegetatio* **97**, 21–38 (1991).
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
- Jürgens, N. Untersuchungen zur Ökologie sukkulenter Pflanzen des südlichen Afrika. *Mitt. Inst. Allg. Bot. Hamburg* **21**, 139–365 (1986).
- Ihlenfeldt, H.-D. Diversification in an arid world: The Mesembryanthemaceae. *Annu. Rev. Ecol. Syst.* **25**, 521–546 (1994).
- Richardson, J. E. et al. Rapid and recent origin of species richness in the Cape flora of South Africa. *Nature* **412**, 181–183 (2001).
- Richardson, J. E., Pennington, R. T., Pennington, T. D. & Hollingsworth, P. M. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science* **293**, 2242–2245 (2001).
- Baldwin, B. G. & Sanderson, M. J. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proc. Natl Acad. Sci. USA* **95**, 9402–9406 (1998).
- Schmiedel, U. & Jürgens, N. Community structure on unusual islands: quartz-fields in the Succulent Karoo, South Africa. *Plant Ecol.* **142**, 57–69 (1999).
- Hendey, Q. B. *Langebaanweg. A Record of Past Life* (South African Museum, Cape Town, 1982).
- Axelrod, D. I. & Raven, P. H. In *Biology and Ecology of Southern Africa* (ed. Werger, M. J. A.) 77–130 (The Hague, Junk, 1978).
- Ward, J. D., Seely, M. K. & Lancaster, N. On the antiquity of the Namib. *S. Afr. J. Sci.* **79**, 175–183 (1983).
- Ward, J. D. & Corbett, I. Towards an age of the Namib. *Transvaal Mus. Monogr.* **7**, 17–26 (1990).
- Magallón, S. & Sanderson, M. J. Absolute diversification rates in Angiosperm clades. *Evolution* **55**, 1762–1780 (2001).
- Eriksson, O. & Bremer, B. Pollination systems, dispersal modes, lifeforms, and diversification rates in Angiosperm families. *Evolution* **46**, 258–266 (1992).
- Mayhew, P. J. Shifts in hexapod diversification and what Haldane could have said. *Proc. R. Soc. Lond. B* **269**, 969–974 (2002).
- Hulbert, R. C. Jr The rise and fall of an adaptive radiation. *Paleobiology* **19**, 216–234 (1993).
- McCune, A. R. In *Molecular Evolution and Adaptive Radiation* (eds Givnish, T. J. & Sytsma, K. J.) 585–610 (Cambridge Univ. Press, 1997).
- Scott, L., Steenkamp, M. & Beaumont, P. B. Palaeoenvironments in South Africa at the Pleistocene–Holocene transition. *Quat. Sci. Rev.* **14**, 937–947 (1995).
- Landrum, J. V. Wide-band tracheids in leaves of genera in Aizoaceae: the systematic occurrence of a novel cell type and its implications for the monophyly of the subfamily Ruschioideae. *Pl. Syst. Evol.* **227**, 49–61 (2001).
- Mauseth, J. D., Uosumi, Y., Plemons, B. J. & Landrum, J. V. Structural and systematic study of an unusual tracheid type in cacti. *J. Plant Res.* **108**, 517–526 (1995).
- Parolin, P. Seed expulsion in fruits of Mesembryanthema (Aizoaceae): a mechanistic approach to study the effect of fruit morphological structures on seed dispersal. *Flora* **196**, 313–322 (2001).
- Klak, C., Khunou, A., Reeves, G. & Hedderson, T. A phylogenetic hypothesis for the Aizoaceae (Caryophyllales) based on four plastid DNA regions. *Am. J. Bot.* **90**, 1433–1445 (2003).
- Swofford, D. L. *PAUP*4.0b2: Phylogenetic Analysis Using Parsimony* (Sinauer Associates, Sunderland, Massachusetts, 1998).
- Fitch, W. M. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.* **20**, 406–416 (1971).
- Felsenstein, J. Confidence limits on phylogenies: an approach using bootstrap. *Evolution* **39**, 783–791 (1985).
- Sanderson, M. J. A non parametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1232 (1997).
- Rambaut, A. & Charleston, M. *TreeEdit version 1.0 alpha 4-61* [online] (<http://evolve.zoo.ox.ac.uk/software/TreeEdit.html>) (2000).

- Scott, L., Anderson, H. M. & Anderson, J. M. In *Vegetation of Southern Africa* (eds Cowling, R. M., Richardson, D. M. & Pierce, S. M.) 62–84 (Cambridge Univ. Press, 1997).
- Tankard, A. J. & Rogers, J. Late Cenozoic palaeoenvironments on the west coast of Southern Africa. *J. Biogeogr.* **5**, 319–337 (1978).
- Wikström, N., Savolainen, V. & Chase, M. Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. Lond. B* **268**, 1–10 (2001).

Acknowledgements We thank P. Bruyns, A. Ellis and T. Verboom for critical discussion and comments, and G. Aguilar, F. Conrad and A. Khunou for technical support. This research was supported by the National Botanical Institute, Cape Town, the Mesemb Study Group (UK) and grants from The National Research Foundation, and by The University of Cape Town's University Research Committee (T.A.H.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.K. (klak@botzoo.uct.ac.za). All sequences have been submitted to EMBL (accession numbers AJ438210–438215, AJ438218–438219, AJ438223–438228, AJ438230, AJ438232, AJ438239, AJ438250, AJ438253, AJ438255, AJ438257–438259, AJ438261, AJ438998–439004, AJ439007–439008, AJ439013–439018, AJ439021, AJ439023, AJ439030, AJ439042, AJ439045, AJ439047, AJ439049–439051, AJ439053, AJ53270–532824, AJ558034–558102 and AJ532594–532606).

A euprimate skull from the early Eocene of China

Xijun Ni¹, Yuanqing Wang¹, Yaoming Hu^{1,2,3} & Chuankui Li¹

¹Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

²Division of Paleontology, American Museum of Natural History, New York 10024, USA

³Biology Program (Ecology, Evolutionary Biology, and Behavior), Graduate School and City College, City University of New York, New York 10016, USA

The debut of undoubted euprimates (primates of modern aspect^{1,2}) was in the early Eocene, about 55 Myr ago. Since their first appearance, the earliest euprimates can be distinguished as *Cantius*, *Donrussellia* and *Teilhardina*^{2–4}. Nonetheless, the earliest euprimates are primarily known from isolated teeth or fragmentary jaws. Here we describe a partially preserved euprimate skull with nearly complete upper and lower dentition, which represents a new species of *Teilhardina* and constitutes the first discovery of the genus in Asia. The new species is from the upper section of Lingcha Formation, Hunan Province, China, with an estimated age of 54.97 Myr ago⁵. Morphology and phylogeny analyses reveal that the new species is the most primitive species of *Teilhardina*, positioned near the root of euprimate radiation. This discovery of the earliest euprimate skull known to date casts new light on the debate^{6–12} concerning the adaptive origin of euprimates, and suggests that the last common ancestor of euprimates was probably a small, diurnal, visually oriented predator.

Primates Linnaeus, 1758
Omomyidae Trouessart, 1879
Teilhardina Simpson, 1940
Teilhardina asiatica sp. nov.

Holotype. A partial skull with associated lower jaws (IVPP V12357, Figs 1, 2).

Included material. An isolated lower incisor (IVPP V12357-4) and two additional partial lower jaws (IVPP V12060, V13762).

Horizon and locality. Upper part of the Lingcha Formation, Hengyang Basin, China; earliest Eocene⁵.

Diagnosis. Differs from *T. belgica* and *T. americana* in having a

less-reduced lower premolar P_1 , which is loosely spaced and aligned in line with P_2 .

The braincase of *T. asiatica* is rounded and large (Fig. 1), in sharp contrast to the low and small braincase in plesiadapiforms, such as *Ignacius*, *Palaechthon* and *Plesiadapis*. The skull roof of the new species is smooth and bears no sagittal crest, as on other Eocene omomyids, such as *Tetanius homunculus* and *Shoshonius cooperi*. The infraorbital foramen is relatively large (1 mm in height, 0.83 mm in breadth) and situated above upper premolar P^2 , indicating a significantly shortened snout. The relative size of this foramen resembles those in insectivores or plesiadapiforms, and is greater than in extant and known fossil euprimates. Animals with greater infraorbital foramina are expected to have better-developed vibrissae¹³. The orbits are relatively well-preserved, although their zygomatic margins are broken and the frontal bone is slightly distorted dorsoventrally. The zygomatic process of the frontal is long, flat and narrow. Its irregular broken surface suggests that a postorbital bar was present. The preserved portion of the orbital margin is a well-defined, smoothly curved edge, demonstrating that the compression of the frontal bone caused little distortion of the orbit.

Relative to the skull length, the orbital size (mean = 6.92 mm) is much larger than those of plesiadapiforms, but is moderate within euprimates. Also different from plesiadapiforms, the orbits of the new species are significantly convergent with narrow interorbital breadth (3.97 mm). Relative to skull length, the interorbital breadth fits to the best-fit line for early Tertiary euprimates¹⁴. The convergence degree (51°) falls in the range of basal euprimates¹⁵ and that of extant prosimians¹⁶. Because the frontal bone is crushed, precise measurement for the frontation of the orbits is not available.

However, because the well-preserved inferior-medial margin of the orbit is inclined caudally, the eyes must have had a low degree of frontation.

The lower jaws are nearly complete. The symphysis is unfused, and nearly horizontal, similar to that of *T. belgica*. Both the coronoid process and condyle are much higher than the tooth row. The angular process is straight and long, with its end posterior to the condyle. *Donrussellia magna*, another known earliest-Eocene eupri-mate with well-preserved mandible¹⁷, differs from *T. asiatica* in possessing a much lower condyle and less prominent angular process.

The dental morphology of *T. asiatica* falls well in the diagnosis of the genus *Teilhardina*. In particular, the new species greatly resembles *T. belgica*.

The upper canine is sharp, slender and bodkin-like. Teeth P^{1-2} are reduced in size and simple in morphology. The P^3-M^3 teeth are closely similar to those of *T. belgica*. The minor differences between the two species mainly lie in molars: the M^1 in *T. asiatica* is more smoothly curved in its anterolingual margin and more protuberant at its posterolingual corner. The upper teeth of *T. americana* are saliently different from those of *T. asiatica* and *T. belgica* in having the *Nannopithec*-fold, a stronger lateral cingulum, and a more rectangular outline in the M^1 and M^2 .

The lower canine is also bodkin-like with slightly procumbent root. The large canine alveolus in the mandible of *T. belgica* indicates the presence of a tall and sharp canine, as in *T. asiatica*. The lower canine of *T. americana* is also tall, but clearly premolariform. Its crown is buccolingually expanded, and has a definite border above

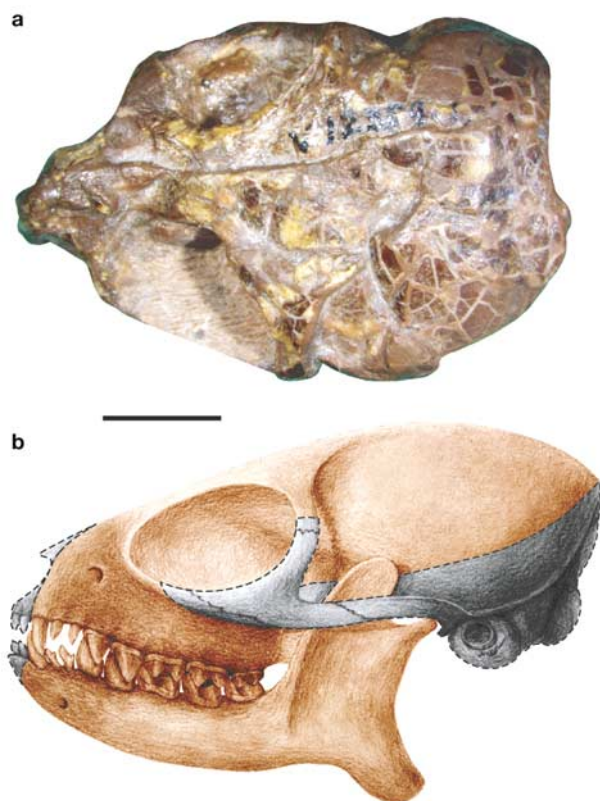


Figure 1 The skull of *Teilhardina asiatica* sp. nov. (IVPP V12357). **a**, Dorsal view of the skull. **b**, Reconstruction of the skull based on IVPP V12357, with grey shadow indicating the missing parts. Scale bar, 5 mm.

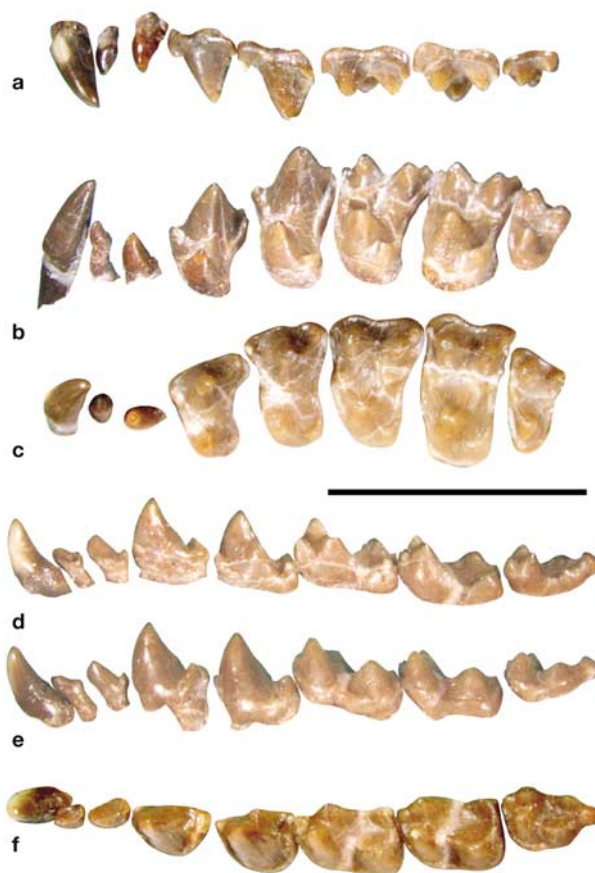


Figure 2 The dentition of *Teilhardina asiatica* sp. nov. (IVPP V12357). **a-c**, Lateral, occlusolingual and occlusal view of the left upper teeth row. **d-f**, Reversed lingual, lateral and occlusal view of the left lower teeth row. Scale bar, 5 mm.

the root. The premolarization of the lower canine is probably associated with the incisor I_1 enlargement. Later North American species of *Teilhardina* have an enlarged I_1 , and their lower canines are premolariform and greatly reduced in size.

The P_{1-2} teeth are small and simple in morphology. Both are loosely spaced and implanted in line with other cheek teeth. Reduction of the P_1 is an evolutionary trend among species of *Teilhardina*¹⁸. *T. belgica* and *T. americana* possess a laterally displaced and very small P_1 , whereas other derived North American species may completely lack the tooth. The P_1 of *T. asiatica* is less reduced and not laterally displaced in comparison with those of *T. belgica* and *T. americana*, representing a more primitive condition within the genus.

The differences in the P_3 – M_3 teeth among *T. belgica*, *T. americana* and *T. asiatica* are minor. *T. americana* is more derived than the other two species in having a relatively higher metaconid on the P_4 and a higher entoconid on lower molars. *T. asiatica* possesses a slightly more distinctive cuspule on the posterolingual end of the heel of P_3 and P_4 .

An isolated incisor derived from the type specimen is identified as I_2 , judging from its wear facet. The tooth is tiny compared with the canine and has a relatively robust root. The neck and root of the tooth are mesiodistally compressed, whereas the crown is labiolingually compressed. The previously known lower incisor of the

North American *Teilhardina* is enlarged and lanceolate. The incisor alveolus of *T. belgica* indicates that the primitive state of the lower incisor of *Teilhardina* may be small. *T. asiatica* shows that the un-enlarged lower incisor of *Teilhardina* is not lanceolate but more nearly spatulate.

Previously, six species of *Teilhardina* have been reported based primarily on dental material: *T. belgica*, *T. americana*, *T. brandti*, *T. crassidens*, *T. tenuicula*, and *T. demissa*. Among those, *T. belgica* is considered to be the most primitive species, whereas *T. americana* is more derived than *T. belgica*. The earliest Eocene *T. brandti* is represented by only one M_2 , and its taxonomic validity is questionable owing to the limited material¹⁹. Other North American species of *Teilhardina* are all considered to be more derived than *T. americana*. *T. asiatica* is morphologically very close to *T. belgica*, and some dental characters, such as less reduced P_1 and loosely spaced anterior premolars, are more primitive than the latter.

Steinius vespertinus was previously considered to be dentally as primitive as *Teilhardina* or more so^{3,20}. However, the anteroposteriorly compacted P_1 and P_2 , enlarged I_1 , wider lateral cingulids on premolars and molars in *S. vespertinus* indicate that it is more derived than *T. asiatica* and *T. belgica*. The relative size of P_3 to M_1 and the length/width ratios of P_3 and P_4 among *S. vespertinus*, *T. asiatica* and *T. belgica* are similar, making it improper to evaluate evolutionary degree among these species using those features. Relatively larger M_3 in *S. vespertinus* is probably a derived character of omomyines rather than a primitive trait relative to that of *Teilhardina*.

Our morphology analysis suggests that *T. asiatica* is a basal omomyid and cannot be far from the euprimate stem. This view is consistent with the results of a cladistic analysis (Fig. 3). Our phylogenetic analysis supports the previous hypothesis that euprimates are basally dichotomous^{4,14,21,22}. *T. asiatica* and *T. belgica* form a sister-group that is situated at the base of the omomyid–tarsier–

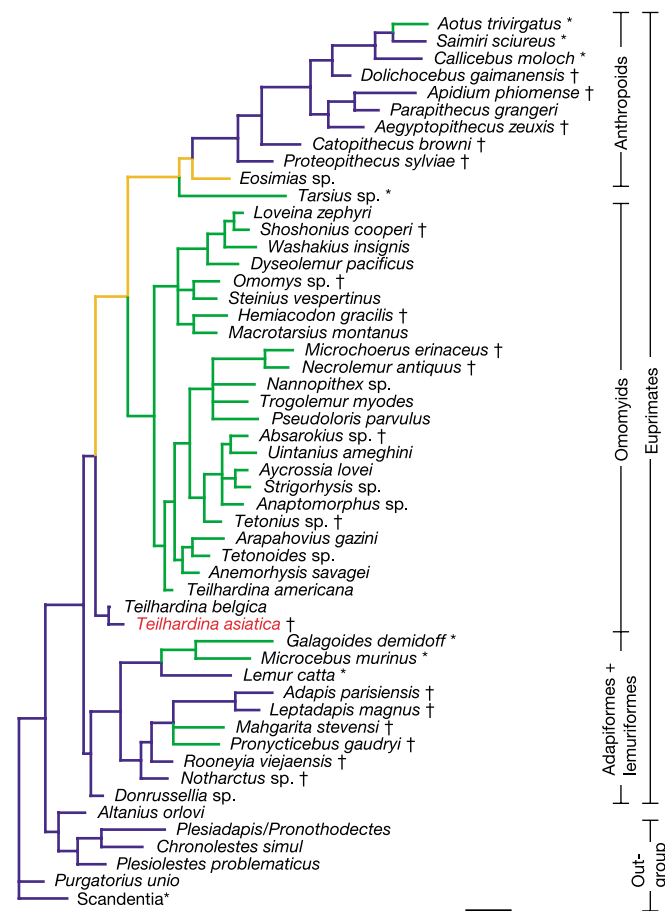


Figure 3 Strict consensus of 33 equally parsimonious trees with the optimization of activity patterns. Tree length = 2,076, Consistency index (CI) = 0.3685, Retention index (RI) = 0.5519. Asterisks denote extant taxa. Daggers denote the terminal taxa presenting reconstructions of activity patterns (data are from refs 7, 23, 30). Blue, diurnal; green, nocturnal; orange, equivocal. Scale bar, 30 characters.

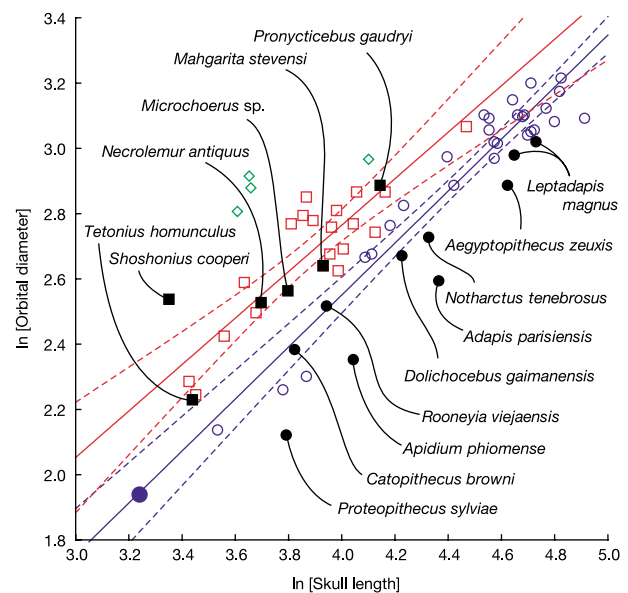


Figure 4 Bivariate plot of $\ln$ [orbital diameter] and $\ln$ [skull length] for extant and fossil primates. The red and blue lines are linear fits for extant nocturnal strepsirrhines and extant diurnal primates respectively. Dashed lines show 95% confidence intervals. Open blue circle, extant diurnal primates; red square, extant nocturnal strepsirrhines; green rhombus, extant nocturnal haplorhines; solid blue circle, *Teilhardina asiatica*; filled black circle, reconstructed diurnal fossil primates; filled black square, reconstructed nocturnal fossil primates. We note that *T. asiatica* falls in the extrapolated range of diurnal primates.

anthropoid clade. Interestingly, however, the more derived *T. americana* is not grouped with *T. asiatica* and *T. belgica*, suggesting that the conventional *Teilhardina* is probably a polyphyletic taxon (see Supplementary Information).

Based on the functional/adaptive interpretation of the relative size of orbits, omomyids, except phylogenetically enigmatic *Rooneyia*, were reconstructed as nocturnal animals^{7,13,23}. *T. asiatica* as the most primitive omomyid, however, shows relatively smaller orbital size compared to other omomyids, and falls in the extrapolated range of diurnal forms rather than nocturnal ones (Fig. 4, see also Supplementary Information). This argues that the nocturnality of omomyids must be a derived activity pattern. Because *T. asiatica* is phylogenetically near the root of the euprimate radiation, the reconstruction of diurnal activity pattern for the species counts as evidence against the proposal that the ancestral euprimates were nocturnal. On the basis of the current phylogenetic relationship (Fig. 3), it is the most parsimonious interpretation that the last common ancestor of euprimates was diurnal.

The body weight of *T. asiatica* estimated from the M_1 size is about 28 g, well under the Kay's threshold^{4,24,25}. The M_2 shear ratio²⁶ of 2.09 falls into the range of primates feeding on insects²⁶. In addition, with its large canines, and sharp-pointed premolars with well-developed shearing crests, *T. asiatica* is undoubtedly insectivorous. Relatively large and convergent orbits indicate that *T. asiatica* must rely more on vision for predation than plesiadapiforms (although it may also have well-developed vibrissae). From this aspect, the new skull undermines a recent hypothesis that the euprimate was evolved from a *Carpolestes*-like terminal fruit-feeder in the latest Palaeocene¹¹. The last common ancestor of stem euprimates, if it mirrored *T. asiatica* in morphology, would be portrayed as a small, diurnal, visually oriented predator. This would have positioned earliest euprimates in a very different ecological niche from that occupied by apparently more frugivorous, less visually oriented carpolestids, and would therefore argue against the hypothesis that euprimates replaced carpolestids ecologically in the Eocene¹¹ (see also ref. 12). □

Methods

Measurements

Measurements were taken with an optical reticle on a Motic microscope. Infraorbital foramen and tooth measurements were taken at magnifications of $\times 30$, others at $\times 6$. Teeth are denoted as I for incisor, C for canine, P for premolar, and M for molar. Super- and subscript numbers indicate upper and lower teeth.

Claustic analysis

The data used in the analyses were derived from a published matrix²¹. We modified the original data set to incorporate recent findings, expanding it to include 12 additional characters and three additional taxa. In total, 303 characters (194 dental, 49 cranial, 56 postcranial and four soft tissue characters) and 52 taxa (five as outgroup) were included. All characters were equally weighted; partial characters were determined as ordered. A heuristic search was undertaken in PAUP 4.0b10 (ref. 27) with 5,000 replications. A strict consensus tree of all equally most parsimonious alternatives was computed.

Dietary habits and activity pattern reconstruction

Primate dietary habits are closely linked with body size. Leaf-eating primates have body weights of greater than 500 g (Kay's threshold), whereas insectivorous primates tend to weigh less than this limit^{4,24,25}. We estimated the body weight of *T. asiatica* from the M_1 area by using Gingerich's empirical equation for tarsoids²⁵. An insectivorous primate can also be distinguished from a fruit-eating one by its more-developed molar shearing crests^{24,26,28}. We measured the length of shearing crests 1–6 on M_2 as defined in ref. 28, and calculated the mean shear ratio following ref. 26.

Osteological evidence from living primates reveals that the orbit size relative to skull length is strongly correlated with activity pattern^{7,13,14,23}. When $\ln[\text{orbit length}]$ is plotted against $\ln[\text{skull length}]$, the nocturnal and diurnal primates have different best-fit lines. The nocturnal primate has relatively larger orbit length than the diurnal form for a given skull length. Natural logarithmic species means of orbital diameter and skull length for 52 extant primate species (28 diurnal primates; 20 nocturnal strepsirrhines; four nocturnal haplorhines) from publications²³ were analysed using a bivariate plot. Large-bodied diurnal primates (for example, baboon, gelada, mandrill, chimpanzee and human) are not included in the analysis because they are much more derived than other primates in skull shape. Regression lines for diurnal primates and nocturnal strepsirrhines were calculated separately: for diurnal primate, $\ln[\text{orbital diameter}] = -0.64 + 0.80 \times \ln[\text{skull length}]$,

$R^2 = 0.92$; for nocturnal strepsirrhines, $\ln[\text{orbital diameter}] = -0.08 + 0.71 \times \ln[\text{skull length}]$, $R^2 = 0.79$. Measurements (in millimetres) for *T. asiatica* and the other 15 fossil euprimates were plotted together with the extant species to determine the most likely activity patterns.

To detect the evolution of euprimates' activity patterns from a phylogenetic perspective, we optimized the reconstructed activity patterns for fossil euprimates together with the extant ones on the phylogeny tree, using the maximum parsimony rule. Character mapping was conducted in MacClade 3.0 (ref. 29).

Received 25 April; accepted 8 October 2003; doi:10.1038/nature02126.

1. Simons, E. L. *Primate Evolution: an Introduction to Man's Place in Nature* (Macmillan, New York, 1972).
2. Szalay, F. S. & Delson, E. *Evolutionary History of the Primates* (Academic, London, 1979).
3. Rose, K. D., Godinot, M. & Bown, T. M. in *Anthropoid Origins* (eds Fleagle, J. G. & Kay, R. F.) 1–28 (Plenum, New York, 1994).
4. Fleagle, J. G. *Primate Adaptation and Evolution* 2nd edn (Academic, London, 1999).
5. Bowen, G. J. et al. Mammalian dispersal at the Paleocene/Eocene boundary. *Science* **295**, 2062–2065 (2002).
6. Cartmill, M. in *The Functional and Evolutionary Biology of Primates* (ed. Tuttle, R.) 97–122 (Aldine Atherton, Chicago, 1972).
7. Heesy, C. P. & Ross, C. F. Evolution of activity patterns and chromatic vision in primates: morphometrics, genetics and cladistics. *J. Hum. Evol.* **40**, 111–149 (2001).
8. Li, W.-H. Genetic systems of color vision in primates. *Am. J. Phys. Anthropol.* **30** (suppl.), 318 (2000).
9. Tan, Y. & Li, W.-H. Trichromatic vision in prosimians. *Nature* **402**, 36 (1999).
10. Surridge, A. K., Osorio, D. & Mundy, N. I. Evolution and selection of trichromatic vision in primates. *Trends Ecol. Evol.* **18**, 198–205 (2003).
11. Bloch, J. I. & Boyer, D. M. Grasping primate origins. *Science* **298**, 1606–1610 (2002).
12. Kirk, E. C., Cartmill, M., Kay, R. F. & Lemelin, P. Comment on "Grasping primate origins". *Science* **300**, 741 (2003).
13. Kay, R. F. & Cartmill, M. Cranial morphology and adaptations of *Palaeothentes* and other Paromomyidae (Plesiadapoidea? Primates), with a description of a new genus and species. *J. Hum. Evol.* **6**, 19–53 (1977).
14. Martin, R. D. *Primate Origins and Evolution: a Phylogenetic Reconstruction* (Chapman and Hall, London, 1990).
15. Ravosa, M. J., Noble, V. E., Hylander, W. L., Johnson, K. R. & Kowalski, E. M. Masticatory stress, orbital orientation and the evolution of the primate postorbital bar. *J. Hum. Evol.* **38**, 667–693 (2000).
16. Ross, C. F. Allometric and functional influences on primate orbit orientation and the origins of the Anthropoidea. *J. Hum. Evol.* **29**, 201–227 (1995).
17. Godinot, M. et al. Nouvelles données sur les mammifères de Palette (Eocène inférieur, Provence). *Münchener Geowiss. Abh.* **10**, 237–288 (1987).
18. Bown, T. M. & Rose, K. D. Patterns of dental evolution in Early Eocene Anaptomorphine primates (Omomyidae) from the Bighorn Basin, Wyoming. *J. Paleontol.* **61**(s; supplement), 1–162 (1987).
19. Gunnell, G. F. & Rose, K. D. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 45–82 (Cambridge Univ. Press, Cambridge, 2002).
20. Rose, K. D. & Bown, T. M. Additional fossil evidence on the differentiation of the earliest euprimates. *Proc. Natl Acad. Sci. USA* **88**, 98–101 (1991).
21. Ross, C., Williams, B. A. & Kay, R. F. Phylogenetic analysis of anthropoid relationships. *J. Hum. Evol.* **35**, 221–306 (1998).
22. Kay, R. F., Ross, C. & Williams, B. A. Anthropoid origins. *Science* **275**, 797–804 (1997).
23. Kay, R. F. & Kirk, E. C. Osteological evidence for the evolution of activity pattern and visual acuity in primates. *Am. J. Phys. Anthropol.* **113**, 235–262 (2000).
24. Kay, R. F. The functional adaptation of primate molar teeth. *Am. J. Phys. Anthropol.* **43**, 195–216 (1975).
25. Gingerich, P. D. Early Cenozoic Omomyidae and the evolutionary history of Tarsiiform primates. *J. Hum. Evol.* **10**, 345–374 (1981).
26. Williams, B. A. & Covert, H. H. New early Eocene anaptomorphine primate (Omomyidae) from the Washakie Basin, Wyoming, with comments on the phylogeny and paleobiology of anaptomorphines. *Am. J. Phys. Anthropol.* **93**, 323–340 (1994).
27. Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4* (Sinauer, Sunderland, 2000).
28. Kay, R. F. & Hiemeke, K. M. Jaw movement and tooth use in recent and fossil primates. *Am. J. Phys. Anthropol.* **40**, 227–256 (1974).
29. Maddison, W. P. & Maddison, D. R. *MacClade: Analysis of Phylogeny and Character Evolution, Version 3.0* (Sinauer, Sunderland, 1992).
30. Covert, H. H. in *Creatures of the Dark: The Nocturnal Prosimians* (eds Alterman, L., Doyle, G. A. & Izard, M. K.) 495–509 (Plenum, New York, 1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Xie for fossil preparation. We are grateful to J. Meng for comments on drafts of this manuscript, R. Kay for providing his published data and for suggestions, and K. C. Beard for discussions. This research is funded by the National Natural Science Foundation of China, the Major Basic Research Project of Ministry of Sciences and Technology, China, and the Chinese Academy of Sciences.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.N. (ni_xj@263.net).

Unsaturated fatty acid content in seston and tropho-dynamic coupling in lakes

Dörthe C. Müller-Navarra^{1*}, Michael T. Brett², Sangkyu Park¹, Sudeep Chandra¹, Ashley P. Ballantyne², Eduardo Zorita³ & Charles R. Goldman¹

¹University of California, Department of Environmental Science and Policy, One Shields Avenue, Davis, California 95616, USA

²University of Washington, Department of Civil & Environmental Engineering, Box 352700, 301 More Hall, University of Washington, Seattle, Washington 98195-2700, USA

³Institute for Coastal Research GKSS Research Centre, Max-Planck-Strasse, D-21502 Geesthacht, Germany

* Present address: Institute for Hydrobiology and Fishery Science, University of Hamburg, Olbersweg 24, D-22767 Hamburg, Germany

Determining the factors that control food web interactions is a key issue in ecology^{1,2}. The empirical relationship between nutrient loading (total phosphorus) and phytoplankton standing stock (chlorophyll *a*) in lakes was described about 30 years ago³ and is central for managing surface water quality. The efficiency with which biomass and energy are transferred through the food web and sustain the production of higher trophic levels (such as fish) declines with nutrient loading and system productivity^{4,5}, but the underlying mechanisms are poorly understood. Here we show that in seston (fine particles in water) during summer, specific ω 3-polyunsaturated fatty acids (ω 3-PUFAs), which are important for zooplankton^{6–10}, are significantly correlated to the trophic status of the lake. The ω 3-PUFAs octadecatetraenoic acid, eicosapentaenoic acid (EPA) and docosahexaenoic acid, but not α -linolenic acid, decrease on a double-logarithmic scale with increasing total phosphorus. By combining the empirical relationship between EPA-to-carbon content and total phosphorus with functional models relating EPA-to-carbon content to the growth and egg production of daphnids⁸, we predict secondary production for this key consumer. Thus, the decreasing efficiency in energy transfer with increasing lake productivity can be explained by differences in ω 3-PUFA-associated food quality at the plant–animal interface.

The efficiency with which biomass and energy are transferred across the plant–animal interface is highly variable^{11,12}, and these trophic levels can even become uncoupled in very productive lakes^{5,13}. This is especially true during the summer season, when food limitation for zooplankton is most prevalent^{14,15} and when the effects of eutrophication are most obvious—for example, noxious phytoplankton blooms are most likely to occur in summer^{15,16}. So far, there is no clear understanding of what determines the transfer efficiency from phytoplankton to zooplankton, which in turn can influence the zooplankton's capacity to suppress phytoplankton biomass. There is, however, a broad consensus that food quality for zooplankton is a pivotal factor^{2,14,17,18}. In addition to differences in digestibility¹⁷, pronounced differences in elemental stoichiometry^{18,19} and the biochemical make-up^{2,6–9} of plants and animals can be key components that determine food quality for zooplankton.

Studies suggest that the highly unsaturated fatty acids (ω 3-HUFAs) EPA and docosahexaenoic acid (DHA) are important biochemical constituents in the natural diet of zooplankton because the prevalence of these fatty acids in seston is a strong predictor of zooplankton growth^{6–8} and may be crucial in shaping food web pyramids¹⁰. To test whether the decrease in trophic interaction strength with increasing nutrient (total phosphorus; TP) loading

can be associated with food quality for the primary producer measured as ω 3-PUFA content, we sampled and analysed seston from lakes differing in food web structure and encompassing the natural range of productivity from ultra-oligotrophic to hyper-eutrophic states.

Several lake productivity measures showed the expected positive trend with increasing lake TP (ref. 3). For example, chlorophyll *a* (Chl-*a*; $\log(\text{Chl-}a) = 1.05 \times \text{TP} + 4.8$; $P < 0.005$; correlation coefficient squared, $r^2 = 0.56$) and particulate organic carbon (POC; $\ln(\text{POC}) = 0.55 \times \text{TP} - 2.25$; $P < 0.01$; $r^2 = 0.45$) increased, whereas water transparency (Secchi depth) decreased with increasing TP in these lakes (see Supplementary Information). Our estimates for the slope and intercept of the relationships agree with previously published values⁵.

In contrast to Chl-*a* and POC as measures of zooplankton food

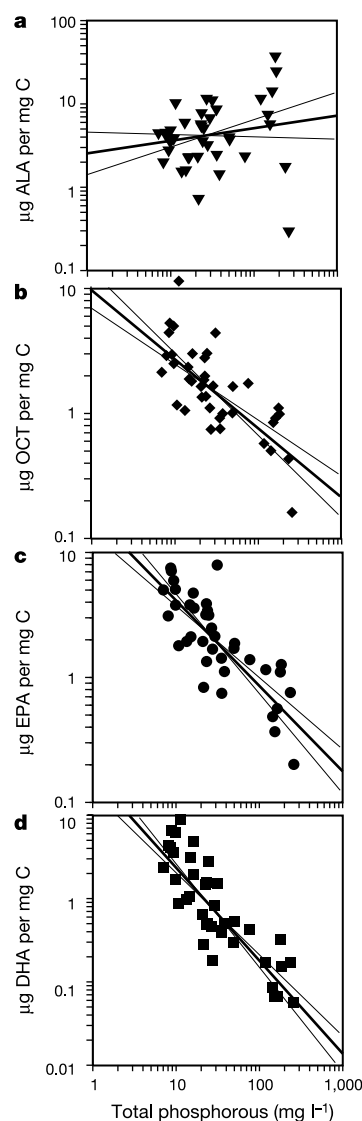


Figure 1 Regressions between lake TP concentration and sestonic ω 3-PUFA-to-C contents. **a**, For ALA (18:3 ω 3), $\ln[\text{ALA} (\mu\text{g per mg C})] = 0.15 \times \ln(\text{TP}) + 0.94$; $P = 0.0531$; $r^2 = 0.03$. **b**, For OCT (18:4 ω 3), $\ln[\text{OCT} (\mu\text{g per mg C})] = -0.69 \times \ln(\text{TP}) + 3.01$; $P < 0.0001$; $r^2 = 0.62$, residual error 0.50. **c**, For EPA (20:5 ω 3), $\ln[\text{EPA} (\mu\text{g per mg C})] = -0.55 \times \ln(\text{TP}) + 2.25$; $P < 0.0001$; $r^2 = 0.55$; residual error 0.47. **d**, For DHA (22:6 ω 3), $\ln[\text{DHA} (\mu\text{g per mg C})] = -1.11 \times \ln(\text{TP}) + 3.42$; $P < 0.0001$; $r^2 = 0.73$; residual error 0.66. Each point represents one observation. Note that the y axis is tenfold larger in **a** than in **b–d**.

quantity, food quality of the summer seston assemblages, measured as the content of the ω 3-HUFAs EPA and DHA relative to the carbon (C) content, was strongly negatively correlated to the trophic status of the lake across the whole naturally occurring gradient of productivity (Figs 1c, d, and 2). This finding can elucidate the observed trend that trophic transfer efficiency and food web strength decrease with increasing trophic status of the lake^{4,5,13}. The border between oligotrophy and eutrophy has been defined³ as 20 μ g of TP per litre, which corresponds to 2 μ g EPA per mg C and about 3 μ g of DHA per mg C (Fig. 2). Notably, a difference between the food quality derived from ω 3-HUFA-to-C content and that derived from the C:P ratio is expected in oligotrophic lakes, where the C:P hypothesis predicts low food quality owing to a high seston C:P ratio under high light-to-nutrient ratios²⁰, and the

sestonic ω 3-HUFA-to-C content predicts high food quality of seston.

The ln–ln negative relationship (that is, the linear decrease on a double logarithmic scale) implies that seston food quality is not distributed equally over the range of TP concentrations (Figs 1 and 2). A decrease in sestonic ω 3-HUFA-associated food quality is very sensitive to TP changes at low TP concentrations, but seems to be almost insensitive at high concentrations of TP. Thus, deviations from the regression curve have a much larger effect at low TP than at high TP. For example, small-scale TP perturbations may result in considerable deviations from the relationship between mean ω 3-HUFA-to-C content and TP at low TP. The degree of non-steady-state between TP and ω 3-HUFA-to-C content at a single time may be inferred from the difference in explained variance of the regressions between single (Fig. 1) and mean (Fig. 2) lake values.

The distribution of ω 3-HUFA-related food quality for zooplankton corroborates observations that responses of phytoplankton to food web alterations (such as different fish stocks) are stronger at low than at high concentrations of TP, and that biomanipulation used as a lake restoration tool is more likely to fail at high TP concentrations²¹, at which ironically it is needed most. By using the empirical relationship between TP and the mean EPA-to-C content in seston (Fig. 2), in conjunction with published zooplankton growth responses to seston EPA-to-C content⁸, we estimated *Daphnia* growth rates and egg production from lake TP concentrations under conditions of non-limiting carbon (Fig. 3). These

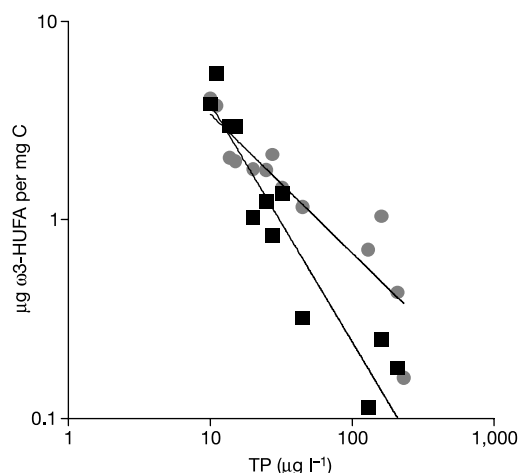


Figure 2 Regressions between lake TP concentration and sestonic EPA and DHA. Grey circles indicate EPA (20:5 ω 3): $\ln[\text{EPA } (\mu\text{g per mg C})] = -0.69 \times \ln(\text{TP}) + 2.78$; $P < 0.0001$; $r^2 = 0.82$; residual error 0.35. Black squares indicate DHA (22:6 ω 3): $\ln[\text{DHA } (\mu\text{g per mg C})] = -1.07 \times \ln(\text{TP}) + 4.4$; $P < 0.0001$; $r^2 = 0.89$; residual error 0.17. Each point represents the mean value of each lake sampled. Given our sample size ($n = 13$), a regression between TP and a dependent variable would have to have a value of $r^2 \geq 0.30$ to be statistically significant at an α -level of 0.005.

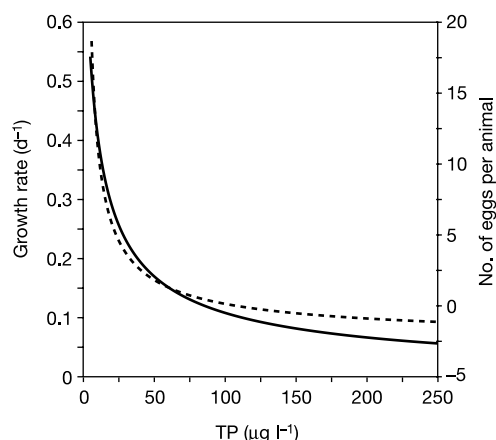


Figure 3 Calculation of the growth rates and egg production of *Daphnia magna*. Growth rates (unbroken line) and egg production (broken line) were calculated from the regression of EPA-to-C content versus TP concentration (Fig. 2), in conjunction with the functional response determined previously (Fig. 3 in ref. 8 for growth rate, $g = 0.74 \times (1 - e^{(-0.25 \times \text{EPA/C} + 0.01)})$, and Fig. 4 in ref. 8 for egg production, eggs = $4 \times \text{EPA/C} - 2.5$).

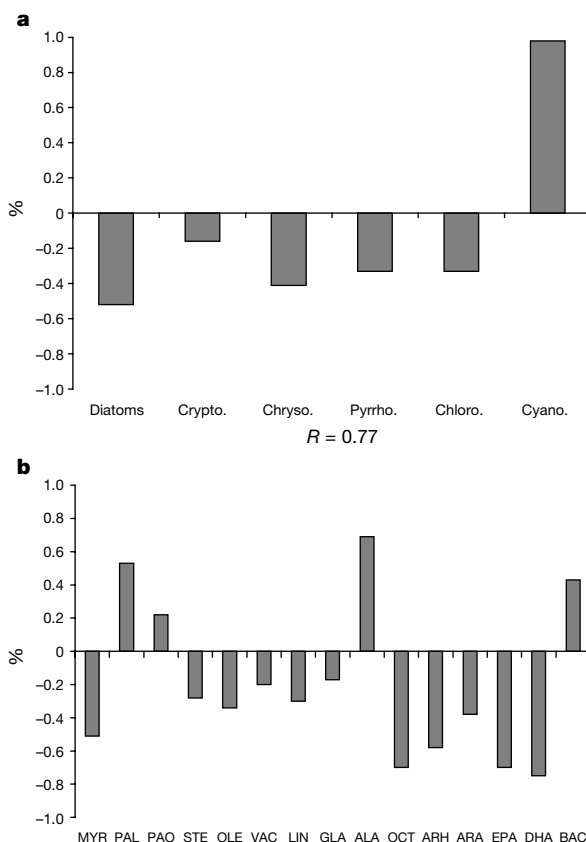


Figure 4 First canonical pattern. **a**, Phytoplankton groups: diatoms, Bacillariophyceae; Crypto., Cryptophyceae; Chryso., Chrysophyceae; Pyrro., Pyrrhophyceae; Chloro., Chlorophyceae; Cyano., Cyanobacteria. **b**, Fatty acids: MYR, 14:0; PAL, 16:0; PAO, 16:1 ω 7; STE, 18:0; OLE, 18:1 ω 9; VAC, 18:1 ω 7; LIN, 18:2 ω 6; GLA, 18:3 ω 6; ARH, 20:0; ARA, 20:4 ω 6; BAC, bacterial fatty acids; other fatty acids are defined in the text. See Methods for an explanation of the statistical analysis.

estimates show that differences in food quality determined by ω 3-HUFA content can have a great effect on the growth rate and egg production of this key grazer. As growth conditions also have an effect on the susceptibility of zooplankton to predation, these effects on food quality might be crucial to promoting grazer populations, thereby influencing the strength of trophic cascades.

By using canonical correlation analysis (CCA), a significantly strong interaction (coefficient of multiple correlation, $R = 0.77$) between phytoplankton group canonical pattern and seston fatty acid canonical pattern was detected, with ω 3-PUFAs dominating the fatty acid pattern (Fig. 4). However, the first canonical pattern explained only 27% of the variance in the phytoplankton data, and only 26% of that in the fatty acid data, suggesting that other factors (such as the physiological state of the phytoplankton and the contribution of detritus and heterotrophic organisms to seston biomass) may have been important in determining the fatty acid composition of seston.

The phytoplankton pattern was dominated by the variation of cyanobacteria, which was in the opposite direction to the variation of all other phytoplankton groups (Fig. 4a). The variation of cyanobacteria covaried with α -linolenic acid (ALA), whereas all other phytoplankton groups covaried with the other ω 3-PUFAs. A high concentration of TP favours cyanobacteria, which can have ALA but have little or none of the other ω 3-PUFAs. In contrast to phytoplankton assemblages dominated by diatoms, chrysophytes and cryptophytes, cyanobacteria can proliferate and build up high standing stocks (blooms) but, owing to low food quality^{22,23}, cannot support higher trophic level (zooplankton) production even at high concentrations of carbon. In addition, as the ω 3-HUFA content of seston is conservatively transferred up the food web into fish, understanding the dynamics of ω 3-HUFA at the base of the food web is a key to their distribution in fish²⁴, which are strongly tied to human nutrition and health²⁵.

The relationships reported here imply that feeding and growth of zooplankton are not independent of, but are coupled to, the nutrient supply to phytoplankton. On the one hand, because of their high ω 3-HUFA content phytoplankton species are more nutritious at low TP and thus can be more easily exploited by zooplankton. On the other hand, phytoplankton species in lakes with high TP will often have a low ω 3-HUFA content and thus are less nutritious, rendering grazers (top-down control) less effective. In contrast to the general view that phytoplankton biomass is mainly controlled from the bottom up, our findings therefore suggest that both bottom-up and top-down processes contribute to the shape of the TP versus Chl-*a* relationship. □

Methods

Sampling

Between mid-June and the beginning of September of 1997–2000, we sampled 13 lakes (1–6 times each) in California, Oregon and Washington, USA (see Supplementary Information). Special attention was paid to ensure that the lakes included various types of lake distributed over a wide gradient of trophic states—which was between 8 and 230 μ g of TP per litre. Both natural and artificial lakes were selected. The lakes sampled were close to a laboratory facility to ensure fast sample processing and proper storage, which are important for the otherwise easily degradable PUFAs. For each lake, we recorded the Secchi depth transparency and collected water samples from the upper mixed layer, generally from a depth of 1 m. For the subalpine lakes (Castle Lake and Lake Tahoe), water was taken from the subsurface to reduce ultraviolet inhibition²⁶. At Lake Tahoe, we collected water from 10 m. At Castle Lake, water was taken from the Chl-*a* maximum (~15 m).

Sample analyses

TP analyses were conducted from raw water samples²⁷. Water for phytoplankton species identification and enumeration was preserved in Lugol's solution, and phytoplankton was counted using an inverted microscope. After filtration onto glass-fibre filters (Whatman GF/C), suspended matter was analysed for Chl-*a*²⁸ with a 10AU fluorometer (Turner), and for POC by either a Model 2400 CHN analyser (Perkin-Elmer) for Stonegate Pond and Lake Tahoe, or a Hydra 20/20 continuous flow isotopic ratio mass spectrometer (PDZ Europa Scientific) for the other lakes. We verified analytical consistency between the latter

two instruments. Fatty acids were analysed after extraction and methylation²⁹ with an HP6890 gas chromatograph (Hewlett Packard) equipped with a programmable temperature vaporizer injector, a fused silica DB-WAX capillary column (J&W Scientific) and a flame ionization detector. ALA, octadecatetraenoic acid (OCT), EPA and DHA are all ω 3-PUFAs, whereas only EPA and DHA are ω 3-HUFAs.

Statistical analyses

The relative abundances of phytoplankton groups and fatty acid compositions were related to each other by the multivariate statistical CCA method³⁰. CCA identifies optimal linearly coupled pattern in two data sets³⁰. Hence for the coupled pattern, the phytoplankton group and fatty acid pattern have the highest possible correlation. In the calculations we used a sample size of 26, which included at least one sample from each lake studied. Data were normalized to their standard deviation. Empirical orthogonal functions (EOFs) were calculated for both the phytoplankton and the fatty acid anomalies (deviation from the mean). The first three EOFs for both the phytoplankton and the fatty acid pattern were used as vectors in the CCA analysis. Explained variability of the phytoplankton EOFs was 28% for the first, 22% for the second and 19% for the third pattern. Thus, 69% of the variability could be captured by the first three EOF functions. Explained variability of the fatty acid EOFs was 29% for the first, 14% for the second and 12% for the third pattern, explaining 55% of the variability with the first three EOF patterns. Both phytoplankton and fatty acid canonical patterns were very similar to their respective EOFs.

Received 18 July; accepted 7 November 2003; doi:10.1038/nature02210.

- Matson, P. A. & Hunter, M. D. (eds) *The Relative Contribution of Top-Down and Bottom-Up Forces in Population and Community Ecology* (Special Feature Ecology 73, Ecological Society of America, 1992).
- Polis, G. A. & Winemiller, K. O. (eds) *Food Webs: Integration of Pattern and Dynamics* (Chapman & Hall, New York, 1996).
- Vollenweider, R. A. Advances in defining the critical loading levels for phosphorus in lake eutrophication. *Mem. Ist. Ital. Idrobiol.* **33**, 53–83 (1976).
- Carpenter, S. R. & Kitchell, J. F. Plankton community structure and limnetic primary production. *Am. Nat.* **124**, 159–172 (1984).
- McQueen, D. J., Post, J. R. & Mills, E. L. Trophic relationships in freshwater pelagic ecosystems. *Can. J. Fish. Aquat. Sci.* **43**, 1571–1581 (1986).
- Jónasdóttir, S. H., Fields, D. & Pantoja, S. Copepod egg production in Long Island Sound USA, as a function of the chemical composition of seston. *Mar. Ecol. Prog. Ser.* **119**, 87–98 (1995).
- Müller-Navarra, D. C. Evidence that a highly unsaturated fatty acid limits *Daphnia* growth in nature. *Arch. Hydrobiol.* **132**, 297–307 (1995).
- Müller-Navarra, D. C., Brett, M. T., Liston, A. M. & Goldman, C. R. A highly unsaturated fatty acid predicts carbon transfer between primary producers and consumers. *Nature* **403**, 74–77 (2000).
- Ravet, J. L., Brett, M. T. & Müller-Navarra, D. C. A test of the role of polyunsaturated fatty acids in phytoplankton food quality for *Daphnia* using liposome supplementation. *Limnol. Oceanogr.* **48**, 1938–1947 (2003).
- Brett, M. T. & Müller-Navarra, D. C. The role of highly unsaturated fatty acids in aquatic food-web processes. *Freshwat. Biol.* **38**, 483–499 (1997).
- Brett, M. T. & Goldman, C. R. Consumer versus resource control in freshwater pelagic food webs. *Science* **275**, 384–386 (1997).
- Micheli, F. Eutrophication, fisheries, and consumer–resource dynamics in marine pelagic ecosystems. *Science* **285**, 1396–1398 (1999).
- McQueen, D. J., Johannes, M. R. S., Post, J. R., Stewart, T. J. & Lean, D. R. S. Bottom-up and top-down impacts on freshwater pelagic community structure. *Ecol. Monogr.* **59**, 289–309 (1989).
- Müller-Navarra, D. C. & Lampert, W. Seasonal patterns of food limitation in *Daphnia galeata*: separating food quantity and quality effects. *J. Plankton Res.* **18**, 1137–1157 (1996).
- Sommer, U., Gliwicz, Z. M., Lampert, W. & Duncan, A. The PEG-model of seasonal succession of planktonic events in fresh waters. *Arch. Hydrobiol.* **106**, 433–471 (1986).
- Schindler, D. W. Detecting ecosystem response to anthropogenic stress. *Can. J. Fish. Aquat. Sci.* **44**, 6–25 (1987).
- DeMott, W. R. & Tessier, A. J. Stoichiometric constraints vs. algal defenses: Testing mechanisms of zooplankton food limitation. *Ecology* **83**, 3426–3433 (2002).
- Elser, J. J., Chrzanowski, T. H., Sterner, R. W. & Mills, K. H. Stoichiometric constraints on food-web dynamics: a whole-lake experiment on the Canadian Shield. *Ecosystems* **1**, 120–136 (1998).
- Sterner, R. W. & Elser, J. J. (eds) *Ecological Stoichiometry. The Biology of Elements from Molecules to the Biosphere* (Princeton Univ. Press, Princeton, NJ, 2003).
- Sterner, R. W., Elser, J. J., Fee, E. J., Guildford, S. J. & Chrzanowski, T. H. The light:nutrient ratio in lakes: the balance of energy and materials affects ecosystem structure and process. *Am. Nat.* **150**, 663–684 (1997).
- Gulati, R. D., Lammens, E. H. R. R., Meier, M.-L. & Van Donk, E. Biomanipulation—tool for water management. *Hydrobiologia* **200/201**, (1990).
- Ahlgren, G. Seasonal variation of fatty acid content in natural phytoplankton in two eutrophic lakes. A factor controlling zooplankton species? *Verh. Int. Verein. Limnol.* **25**, 144–149 (1993).
- Ahlgren, G., Goedkoop, W., Markensten, H., Sonesten, L. & Boberg, M. Seasonal variations in food quality for pelagic and benthic invertebrates in Lake Erken—the role of fatty acids. *Freshwat. Biol.* **38**, 555–570 (1997).
- Ahlgren, G., Sonesten, L., Boberg, M. & Gustafsson, I.-B. Fatty acid content of some freshwater fish in lakes of different trophic levels—a bottom-up effect? *Ecol. Freshwat. Fish* **5**, 15–27 (1996).
- Arts, M., Ackman, R. G. & Holub, B. J. Essential fatty acids in aquatic ecosystems: a crucial link between diet and human health and evolution. *Can. J. Fish. Aquat. Sci.* **58**, 122–137 (2001).
- Wang, K. S. & Chai, T. Reduction in ω 3 fatty acids by UV-B irradiation in microalgae. *J. Appl. Phycol.* **6**, 415–421 (1994).
- Greenberg, A. E., Trussell, R. R., Clesceri, L. S. & Franson, M. A. H. (eds) *Standards and Methods for the Examination of Waters and Wastewaters* (United Book, Baltimore, 1995).
- Marker, A. F., Crowther, C. A. & Gunn, R. J. M. Methanol and acetone as solvents for estimating chlorophyll-*a*, and phaeopigments by spectrophotometry. *Arch. Hydrobiol. Beih. Ergebn. Limnol.* **14**, 52–69 (1980).

29. Kattner, G. & Fricke, H. S. G. Simple gas-liquid chromatographic method for the simultaneous determination of fatty acids and alcohols in wax esters of marine organisms. *J. Chromatogr.* **361**, 263–268 (1986).
30. Mardia, K. V., Kent, J. T. & Bibby, J. M. (eds) *Multivariate Analysis* (Academic, London, 1978).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Liston for assistance; U. Müller, D. Hunter and A. Makulla for phytoplankton counts; and G. Malyj for editing the manuscript. This study was funded by a US National Science Foundation grant to M.T.B., C.R.G. and D.C.M.-N.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.C.M.-N. (doerthe.mueller-navarra@uni-hamburg.de).

SOS response promotes horizontal dissemination of antibiotic resistance genes

John W. Beaber*, Bianca Hochhut* & Matthew K. Waldor

Department of Microbiology, Tufts University School of Medicine and Howard Hughes Medical Institute, 136 Harrison Avenue, Boston, Massachusetts 02111, USA

* These authors contributed equally to this work

Mobile genetic elements have a crucial role in spreading antibiotic resistance genes among bacterial populations. Environmental and genetic factors that regulate conjugative transfer of antibiotic resistance genes in bacterial populations are largely unknown¹. Integrating conjugative elements (ICEs) are a diverse group of mobile elements that are transferred by means of cell-cell contact and integrate into the chromosome of the new host². SXT is a ~100-kilobase ICE derived from *Vibrio cholerae* that encodes genes that confer resistance to chloramphenicol, sulphamethoxazole, trimethoprim and streptomycin³. SXT-related elements were not detected in *V. cholerae* before 1993 but are now present in almost all clinical *V. cholerae* isolates from Asia⁴. ICEs related to SXT are also present in several other bacterial species and encode a variety of antibiotic and heavy metal resistance genes^{4–7}. Here we show that SetR, an SXT encoded repressor, represses the expression of activators of SXT transfer. The 'SOS response' to DNA damage alleviates this repression, increasing the expression of genes necessary for SXT transfer and hence the frequency of transfer. SOS is induced by a variety of environmental factors and antibiotics, for example ciprofloxacin, and we show that ciprofloxacin induces SXT transfer as well.

Table 1 Mitomycin C activates expression of SXT conjugation-associated loci

Site of fusion (background)	β-Galactosidase activity	
	Without mitomycin C	With mitomycin C
Δ <i>setCD::lacZ</i>	15	171
Δ <i>setCD::lacZ</i> (Δ <i>setR</i>)	1,870	1,710
Δ <i>s003::lacZ</i>	11	73
Δ <i>traG::lacZ</i>	17	70
Δ <i>fliR::lacZ</i>	34	33
Δ <i>setCD::lacZ</i> (<i>setR</i> ^{G94E})	112	124

All strains were derivatives of *E. coli* MG1655 harbouring SXT. Mutations in *setR* are indicated in parentheses. The values presented are the means of at least three experiments; standard deviations were less than 10%.

Thus, we present a mechanism by which therapeutic agents can promote the spread of antibiotic resistance genes.

SXT transfer requires *recA* in donor cells, but the molecular basis for this requirement was unclear³. We identified genes at the 3' end of the integrated SXT that regulate SXT transfer⁸. Two loci, *setC* and *setD*, encode transcriptional activators required for SXT excision and transfer⁸. Overexpression of these activators was toxic to cells that harboured SXT but not in cells lacking SXT. *setR*, the 3'-most gene in integrated SXT, is similar to the λ bacteriophage CI repressor and, like this repressor, is predicted to contain both a helix–turn–helix DNA-binding motif and a protease motif (see Supplementary Information). *setR* cannot be deleted from SXT unless the mutation is complemented *in trans* from a plasmid⁸, indicating that the removal of SetR repression of some SXT-encoded factor(s) might be deleterious to cell growth. Because the overexpression of *setC* and *setD* is toxic, we hypothesized that SetR represses these SXT transcriptional activators; we found that *setR* could be deleted in a *setCD* deletion strain.

To measure *setC* and *setD* gene expression, we replaced *setC* and *setD* with a promoterless *lacZ* reporter gene (*setCD::lacZ*). β-Galactosidase activity from this reporter was relatively low (15 Miller units) in the wild-type background; in contrast, the activity was 1,870 Miller units in the *setR* deletion background (Table 1, rows 1 and 2). Introduction of a plasmid carrying *setR* restored the repression of *setCD::lacZ* in the *setR* deletion background (data not shown), confirming that SetR represses *setC* and *setD* expression.

The similarity of SetR to the λ-phage repressor CI suggested that the regulation of SXT transfer might be similar to the regulation of λ lysogeny. In λ lysogens, CI represses prophage gene expression. After DNA damage and the induction of the SOS response, the co-protease activity of RecA is stimulated and promotes the autolysis of CI, alleviating CI-mediated repression and beginning the phage lytic cycle⁹. Because SXT transfer requires *recA* in donor cells³ and SetR is similar to CI, triggering the SOS response might result in a RecA-dependent cleavage and inactivation of SetR, increasing *setC* and *setD* expression and enhancing

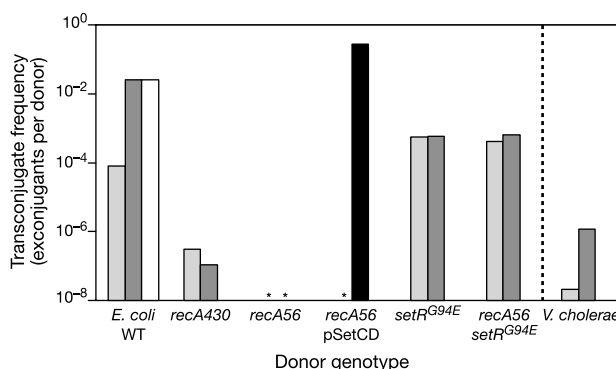


Figure 1 SOS-inducing agents activate SXT transfer. Transconjugate frequency was calculated as transconjugants observed per donor cell as described³. The wild type (WT) and *setR*^{G94E} donor strains are derivatives of *E. coli* strain BW25113 (MG1655 *lacI*^q *rmb*_{T14} Δ*lacZ*_{WJ16} *hsdR*₅₁₄ Δ*araBAD*_{AH33} Δ*rhaBAD*_{LD78})²⁰ containing SXT. The *recA56* and *recA430* donor strains are derivatives of MG1655. *V. cholerae* transfer experiments used MO10 as a donor³. In all cases the recipient was CAG18439 (ref. 21). pSetDC contains *setDC* under the control of an arabinose-inducible promoter⁸. The different growth conditions used for the donor cells before their use in the conjugation assay are represented by light grey bars for growth in LB, dark grey bars for growth in LB containing mitomycin C, white bars for growth in LB containing ciprofloxacin, and black bars for growth in LB containing arabinose. Asterisks represent data that were below the detection limits of the assay (~10⁻⁸). Values presented are the averages of three independent assays; standard deviations were less than 30%.

Induction of the SOS response markedly enhanced SXT transfer. Growth of SXT-containing *Escherichia coli* donor cells in sublethal concentrations of mitomycin C, a DNA-damaging agent that stimulates the SOS response¹⁰, augmented the transfer frequency more than 300-fold (Fig. 1). Similar results were observed with the related element R391 (data not shown). Growing donor cells in ciprofloxacin, a widely used fluoroquinolone antibiotic that activates the SOS response¹¹, also increased SXT transfer (Fig. 1). *V. cholerae* donor cells had a similar response to mitomycin C (Fig. 1) and—in matings of *V. cholerae* to *V. cholerae*—ciprofloxacin (data not shown). To confirm that the stimulation of SXT transfer by mitomycin C was a consequence of SOS induction, we treated *recA* *E. coli* SXT donors with mitomycin C. Both a point mutation that renders RecA co-protease-deficient (*recA430*)¹² and a *recA*-null allele (*recA56*) prevented the induction of SXT transfer by mitomycin C (Fig. 1). Thus, activation of the SOS response in both *E. coli* and *V. cholerae* greatly stimulates the transfer of SXT and SXT-related elements.

Mitomycin C and the SOS response induced expression of genes required for SXT transfer. *lacZ* reporters for *s003* (a gene in the same operon as the SXT integrase gene), *traG* and *setDC* were all induced 4–11-fold by mitomycin C (Table 1), whereas *floR*, the gene mediating chloramphenicol resistance, was not induced, indicating that induction is specific to transfer-associated loci. Mitomycin C did not augment expression from these reporters in *recA*-null cells, indicating that induction of SXT gene expression depends on the SOS response (data not shown). These findings suggest that the SOS response stimulates SXT transfer by increasing the transcription of SXT genes required for the element's transfer.

To assess whether SetR cleavage and inactivation were required for SXT's response to mitomycin C, we generated a SetR variant expected to be resistant to cleavage by RecA. The substitution of glutamate for glycine at the Ala-Gly RecA cleavage site in the CI phage repressor resulted in a non-cleavable repressor and a prophage that was no longer inducible by SOS (ref. 13). Because SetR contains the same putative Ala-Gly cleavage site as CI⁴³⁴ (see Supplementary Information), we introduced a G94E substitution into SetR. We hypothesized that a non-cleavable SetR would prevent the activation of SXT transfer by mitomycin C, because the repression of *setD* and *setC* by SetR would not be relieved after treatment with mitomycin C. In fact, mitomycin C did not increase SXT transfer in the strain harbouring *setR*^{G94E} (Fig. 1), suggesting that SetR inactivation mediates the SOS enhancement of SXT transfer.

The frequency of transfer of the *setR*^{G94E} SXT in a *recA*-null background was nearly identical to that in a wild-type (*recA*⁺) background (Fig. 1). This is in marked contrast to the transfer frequency of the wild-type *setR* SXT in a *recA* background (Fig. 1), indicating that the *setR*^{G94E} mutation is epistatic to the *recA56* mutation. This result places *SetR* downstream of *RecA* in the regulatory circuit controlling SXT transfer.

a

Genetic organization of the *set* operon. The operon consists of the *setC*, *setD*, *s082*, *s083*, *s084*, *s086*, and *setR* genes, followed by the *P_L* promoter. The genes are represented by black arrows pointing right, and the promoter is represented by a black arrow pointing left.

b

β-Galactosidase activity (Miller units) of various strains. The y-axis ranges from 0 to 1,200. The x-axis shows the following strains: SXT⁻ strain, WT SXT, SXT Δ *setCD*, SXT Δ *setCD* Δ *setR*, and SXT *setR*^{G94E}.

Strain	β-Galactosidase activity (Miller units)
SXT ⁻ strain	~1,050
WT SXT	~1,200
SXT Δ <i>setCD</i>	~100
SXT Δ <i>setCD</i> Δ <i>setR</i>	~100
SXT <i>setR</i> ^{G94E}	~100

The diagram illustrates the SetC/SetD quorum sensing system. At the top, a circle labeled **RecA** is shown. An arrow labeled **SOS stimulus** points to a hexagon labeled **RecA***. Below this, a rectangle labeled **Repressor (SetR)** is shown. An arrow labeled **Cleavage** points from the **Repressor (SetR)** to a broken rectangle labeled **Repressor (SetR) inactive**. Below the **Repressor (SetR)**, a T-bar symbol indicates inhibition of the **P_L** promoter. The **P_L** promoter is represented by a bent arrow pointing to the **setR** gene. The **setR** gene is part of a larger DNA sequence that includes the **int** gene, **tra** genes, **setC** gene, **setD** gene, and **s086** gene. The **setC** and **setD** genes are shown as arrows pointing right, while the **setR** gene is shown as a bent arrow pointing left. Below the **setC** and **setD** genes, a box labeled **Activator (SetC and SetD)** has arrows pointing to both genes. The **setR** gene is shown with a T-bar symbol indicating inhibition of the **P_L** promoter.

NATURE | VOL 427 | 1 JANUARY 2004 | www.nature.com/nature

investigated as a potential target of SetR repression because it is upstream of a putative operon that extends from *s086* to *setD* and *setC* and because prokaryotic repressors often act at divergently transcribed promoters. Expression from the promoter upstream of *s086* (designated P_L) was high in a strain lacking SXT (Fig. 2b), but was decreased 13-fold in a strain containing SXT, indicating that P_L is repressed by an SXT-encoded factor (Fig. 2b). *setDC* did not influence expression from P_L in either the presence or the absence of mitomycin C. In contrast, SetR accounted for this repression at P_L because the β -galactosidase activity of the P_L reporter was virtually the same in the $\Delta setR$ and SXT⁻ backgrounds (Fig. 2b). As with the *setCD::lacZ* fusion, treatment with mitomycin C alleviated the SetR repression of P_L only partly, indicating that only a fraction of the total cellular SetR might be inactivated during the SOS response. In the *setR*^{G94E} background P_L was not induced by mitomycin C. SetR repression at P_L seems to be direct because purified epitope-tagged SetR bound to a probe encompassing the region between *s086* and *setR*, which includes P_L (see Supplementary Information).

SXT has co-opted a global cellular response to DNA damage, the SOS response, to control the activity of the SXT repressor, SetR (Fig. 3). SOS might promote the inactivation of SetR by stimulating its autocleavage. Inactivation of SetR relieves the repression of *setC* and *setD*, the transcriptional activators of both the SXT conjugative transfer and integrase genes. Because SetR and the other genes in the 3' regulatory region of SXT seem to be conserved in related elements^{7,14,15}, this regulatory network probably governs the conjugative transfer of all SXT-related elements. Environmental stimuli such as ultraviolet radiation trigger the SOS response and might induce the transfer of SXT. In addition, the SOS response is activated by at least two classes of antibiotics, fluoroquinolones (such as ciprofloxacin) and dihydrofolate reductase inhibitors (such as trimethoprim). Because SOS enhances the conjugative transfer of SXT, the use of certain antimicrobial agents, either clinically or in agricultural settings, might potentiate the horizontal dissemination of antibiotic resistance genes to a broad range of bacterial species. All of these stimuli could account for the rapid manner in which SXT and related elements have spread. □

Methods

Bacterial strains and culture conditions

All bacteria were cultured in Luria–Bertani broth (LB) at 37 °C. Antibiotics were used at the following concentrations: ampicillin, 100 $\mu\text{g ml}^{-1}$; chloramphenicol, 20 $\mu\text{g ml}^{-1}$; ciprofloxacin, 10 ng ml^{-1} (*E. coli*) or 1 ng ml^{-1} (*V. cholerae*); mitomycin C, 200 ng ml^{-1} (*E. coli*) or 20 ng ml^{-1} (*V. cholerae*); sulphamethoxazole, 160 $\mu\text{g ml}^{-1}$; trimethoprim, 32 $\mu\text{g ml}^{-1}$; tetracycline, 10 $\mu\text{g ml}^{-1}$ (*E. coli*) or 1 $\mu\text{g ml}^{-1}$ (*V. cholerae*). Arabinose was added to a final concentration of 0.02%. HW220 *setR*^{G94E} was constructed by allelic exchange by using pOrfRGE as described⁸.

Molecular biology procedures and plasmid construction

Plasmid DNA was prepared by using the Qiaprep Spin Miniprep Kit and Qiaprep Miniprep Kit (Qiagen). Recombinant DNA manipulations were performed by standard procedures¹⁶. The TA Cloning Kit and pBAD TOPO TA Cloning Kit (Invitrogen) were used to clone PCR products. pRep6, which encodes a carboxy-terminally His₆-tagged SetR (SetR-H₆), was constructed by cloning *setR* without a stop codon into pBAD-TOPO. pPs086 was constructed by cloning a PCR fragment of the 250-base-pair region between *s086* and *setR* into pCRII. The fragment was then subcloned into pCB182 (ref. 17). pOrfRGE was made by first amplifying *setR* with primers RepA1 (5'-AAACTTTATCC GAACGACT-3') and RepA2 (5'-CCAGAAATCGATGATAGCTTG-3'). The resulting fragment was cloned into pBAD TOPO and the G94E mutation was introduced by using the Quick Change kit and primers RepGE1 (5'-CTGGGTTTCAGCCGAGGATTGGACT GAAATAGCGG-3') and RepGE2 (5'-CCGCTATTTCAGTCCAATCCTCGGCCTGAA CCCAG-3'). The *setR*^{G94E} was then subcloned into pWM91 (ref. 18), yielding pOrfRGE.

β -Galactosidase assays

Overnight cultures of cells containing either chromosomal or plasmid-borne β -galactosidase fusions were diluted 1:100 into LB containing antibiotics to select for plasmid maintenance and grown for 2 h. Mitomycin C, ciprofloxacin or arabinose was then added to half of the culture and the cells were grown for a further 2 h. β -Galactosidase activity, reported in Miller units, was measured as described¹⁹.

Bacterial matings

Conjugation experiments were conducted as described previously³. In brief, overnight

cultures of differentially marked donor and recipient cells were diluted 1:100 into fresh medium and grown separately for 2 h at 37 °C. Mitomycin C or ciprofloxacin was added as appropriate and cells were grown for a further 1 h at 37 °C. Equal volumes of donor cells and recipient cells were then mixed and spread on a filter on a LB plate. *E. coli* matings were incubated for 1 h at 37 °C; matings of *V. cholerae* to *V. cholerae* were incubated for 16 h at 37 °C. Afterwards, cells were resuspended in LB and dilutions were plated on selective media to enumerate donors, recipients and transconjugants. SXT transfer frequency was calculated as transconjugants per donor cell.

Gel-shift assays

SetR-H₆ was affinity-purified on Ni²⁺-nitrilotriacetate resin (Qiagen) from lysates of *E. coli* LMG194 (Invitrogen) pRep6 grown in the presence of 0.02% arabinose in accordance with the manufacturer's protocol. Probes for gel-shift experiments were made by labelling restriction fragments with ³²P by using the Klenow subunit of DNA polymerase and then gel-purifying them from a 6% retardation gel (Invitrogen). Gel-shift reactions (20 μl volume) were performed by incubating 2,000 c.p.m. of each probe with decreasing amounts (2.5, 1.7, 0.63, 0.25 and 0 ng) of purified SetR-H₆ in a reaction buffer consisting of 80 mM NaCl, 25 mM Tris-HCl pH 8.0, 0.12 mM EDTA, 2.0 mM dithiothreitol, 50 $\mu\text{g ml}^{-1}$ BSA and 25 $\mu\text{g ml}^{-1}$ sonicated salmon-sperm DNA at 4 °C for 1 h. Reactions were analysed on a 6% DNA retardation gel.

Received 12 August; accepted 12 November 2003; doi:10.1038/nature02241.

Published online 21 December 2003.

- Zatyka, M. & Thomas, C. M. Control of genes for conjugative transfer of plasmids and other mobile elements. *FEMS Microbiol. Rev.* **21**, 291–319 (1998).
- Burrus, V., Pavlovic, G., Decaris, B. & Guedon, G. Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* **46**, 601–610 (2002).
- Waldor, M. K., Tschape, H. & Mekalanos, J. J. A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* 0139. *J. Bacteriol.* **178**, 4157–4165 (1996).
- Hochhut, B. *et al.* Molecular analysis of the antibiotic resistance gene clusters in the *Vibrio cholerae* O139 and O1 SXT constains. *Antimicrob. Agents Chemother.* **45**, 2991–3000 (2001).
- Coetzee, J. N., Datta, N. & Hedges, R. W. R factors from *Proteus rettgeri*. *J. Gen. Micro.* **72**, 543–552 (1972).
- Hochhut, B., Beaber, J. W., Woodgate, R. & Waldor, M. K. Formation of chromosomal tandem arrays of the SXT element and R391, two conjugative chromosomally integrating elements that share an attachment site. *J. Bacteriol.* **183**, 1124–1132 (2001).
- Beaber, J. W., Burrus, V., Hochhut, B. & Waldor, M. K. Comparison of SXT and R391, two conjugative integrating elements: definition of a genetic backbone for the mobilization of resistance determinants. *Cell Mol. Life Sci.* **59**, 2065–2070 (2002).
- Beaber, J. W., Hochhut, B. & Waldor, M. K. Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*. *J. Bacteriol.* **184**, 4259–4269 (2002).
- Ptashne, M. *A Genetic Switch* (Blackwell Science & Cell Press, Cambridge, Massachusetts, 1992).
- Little, J. W. In *Regulation of Gene Expression* (eds Lin, E. C. C. & Lynch, A. S.) 453–478 (R. G. Landes, New York, 1996).
- Phillips, L., Culebras, E., Moreno, F. & Baquero, F. Induction of the SOS response by new 4-quinolones. *J. Antimicrob. Chemother.* **20**, 631–638 (1987).
- Walker, G. C. In *Escherichia coli and Salmonella Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 1400–1416 (ASM Press, Washington, DC, 1996).
- Gimble, F. S. & Sauer, R. T. Mutations in bacteriophage λ repressor that prevent RecA-mediated cleavage. *J. Bacteriol.* **162**, 147–154 (1985).
- Boltner, D., MacMahon, C., Pembroke, J. T., Strike, P. & Osborn, A. M. R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements. *J. Bacteriol.* **184**, 5158–5169 (2002).
- Peters, S. E., Hobman, J. L., Strike, P. & Ritchie, D. A. Novel mercury resistance determinants carried by IncJ plasmids pMERPH and R391. *Mol. Gen. Genet.* **228**, 294–299 (1991).
- Ausubel, F. M., *et al.* *Current Protocols in Molecular Biology* (Greene Publishing & Wiley Interscience, New York, 1990).
- Schneider, K. & Beck, C. F. Promoter-probe vectors for the analysis of divergently arranged promoters. *Gene* **42**, 37–48 (1986).
- Metcalf, W. W. *et al.* Conditionally replicative and conjugative plasmids carrying *lacZ* for cloning, mutagenesis, and allele replacement in bacteria. *Plasmid* **35**, 1–13 (1996).
- Miller, J. H. *A Short Course in Bacterial Genetics* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).
- Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
- Singer, M. *et al.* A collection of strains containing genetically linked alternating antibiotic resistance elements for genetic mapping of *Escherichia coli*. *Microbiol. Rev.* **53**, 1–24 (1989).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank B. Davis, S. Chiang, A. Kane and M. Shonn for critical reading of the manuscript. We thank the New England Medical Center GRASP Center for the preparation of media and SetR-H₆. This work was supported in part by funds from a NIH training grant (to J.W.B.), the DFG (to B.H.), the NIH, the Howard Hughes Medical Institute (to M.K.W.) and a pilot project grant from the NEMC GRASP Center.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K.W. (matthew.waldor@tufts.edu).

Defects in RGS9 or its anchor protein R9AP in patients with slow photoreceptor deactivation

Koji M. Nishiguchi¹, Michael A. Sandberg², Aart C. Kooijman⁴, Kirill A. Martemyanov³, Jan W. R. Pott⁴, Stephanie A. Hagstrom^{1*}, Vadim Y. Arshavsky³, Eliot L. Berson² & Thaddeus P. Dryja¹

¹Ocular Molecular Genetics Institute, ²Berman-Gund Laboratory for the Study of Retinal Degenerations, and ³Howe Laboratory of Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts 02114, USA

⁴Department of Ophthalmology, University Hospital Groningen, 9700 RB Groningen, The Netherlands

* Present address: Cole Eye Institute, Cleveland Clinic Foundation, Cleveland, Ohio 44195, USA

The RGS proteins are GTPase activating proteins that accelerate the deactivation of G proteins in a variety of signalling pathways in eukaryotes^{1–6}. RGS9 deactivates the G proteins (transducins) in the rod and cone phototransduction cascades^{7,8}. It is anchored to photoreceptor membranes by the transmembrane protein R9AP (RGS9 anchor protein), which enhances RGS9 activity up to 70-fold^{9–11}. If RGS9 is absent or unable to interact with R9AP, there is a substantial delay in the recovery from light responses in mice^{4,12,13}. We identified five unrelated patients with recessive mutations in the genes encoding either RGS9 or R9AP who reported difficulty adapting to sudden changes in luminance levels mediated by cones. Standard visual acuity was normal to moderately subnormal, but the ability to see moving objects, especially with low-contrast, was severely reduced despite full visual fields; we have termed this condition bradyopsia. To our knowledge, these patients represent the first identified humans with a phenotype associated with reduced RGS activity in any organ.

In transgenic mice lacking *Rgs9*, single-cell recordings of rods in response to light flashes have shown a normal initial rise in the cell membrane potential but an abnormally slowed response recovery⁴. Electroretinograms (ERGs) have revealed that although the response to an initial flash was normal, the response to a subsequent flash separated by 0.03 to 30 s was reduced¹². The mice had no retinal degeneration. These findings prompted us to evaluate the *RGS9*^{14,15} and *R9AP*⁹ genes encoding RGS9 and its anchor protein, respectively, in five unrelated patients having abnormal photo-response recovery. Four of these patients were from the Netherlands and one was from Guatemala. Three of the four unrelated Dutch patients were reported previously (patients 2, 3 and 4 in ref. 16) as having a novel retinal phenotype described in more detail below; the fourth Dutch index patient and the Guatemalan patient have not been previously reported. We found that the four Dutch index patients were homozygous for the missense mutation W299R (TGG to CGG; c.895T → C) in the *RGS9* gene (see Supplementary Information). These patients had no detected mutation in *R9AP*.

Additional evidence indicated that the *RGS9* mutation was responsible for the phenotype in the Dutch patients. First, the segregation of the W299R allele in the four families was consistent with its pathogenicity. In total, ten relatives were evaluated: one affected sibling was a homozygote (patient 1 in ref. 16), and one unaffected sibling, six parents and two children were all heterozygotes (see Supplementary Information for sample pedigrees). Second, the electrophysiological findings in the W299R homozygotes (see below) are similar to those in *rgs9* knockout mice¹². Third, W299 is one of the most highly conserved residues in the catalytic RGS domain among RGS family members in species ranging from *Caenorhabditis elegans* to man. The crystal structure of the RGS

homology domain of RGS9¹⁷ provides an insight into how the W299R mutation could impair the RGS9 function. As illustrated in Fig. 1a, the RGS9 homology domain is composed of nine α -helices, four of which (at the bottom of the figure) directly interact with G proteins, whereas the other five (at the top of the figure) maintain the integrity of the domain by forming a core of hydrophobic interactions. This core has several highly conservative hydrophobic residues (highlighted in the figure) including W299. The substitution of W299 by a positively charged arginine residue would be expected to disrupt these hydrophobic interactions. In order to test this predicted disruption, we expressed the domain with the W299R mutation and evaluated its function. The mutant was much less soluble than the corresponding wild-type domain (not shown), and its ability to stimulate the GTPase activity of transducin was ~20-fold lower than that of the wild-type domain. (The ratio of the wild-type/mutant k_{GAP} was 19.5 ± 1.5 (mean $\pm$

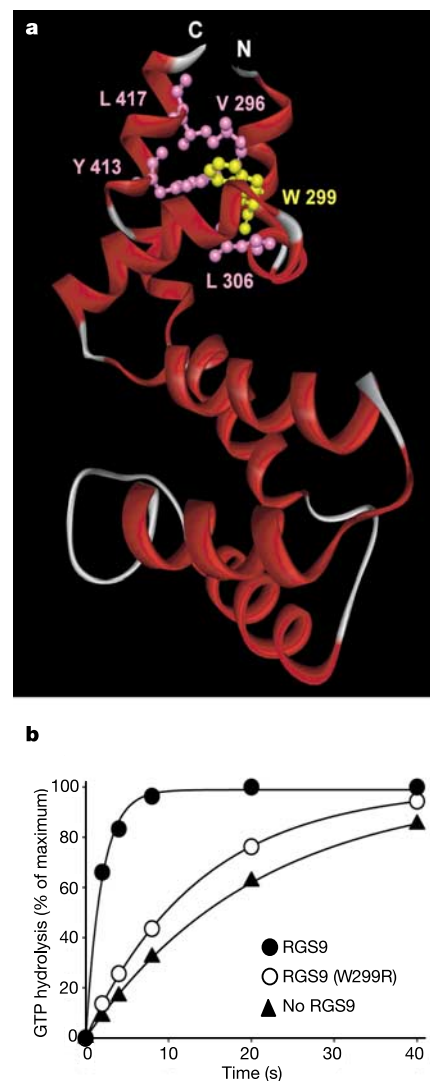


Figure 1 W299 in the RGS9 catalytic domain. **a**, W299 (yellow) and other residues in the hydrophobic core (V296, L306, Y413 and L417—pink) of the RGS9 catalytic domain, based on published coordinates¹⁷. Red, α -helices. **b**, The activation of bovine rod transducin GTPase by the catalytic domains (residues 287–434) of wild-type RGS9 and the W299R mutant. k_{GAP} (the difference between the rate of transducin GTPase activity in the presence of RGS9 and its basal GTPase activity) was $0.45 \pm 0.04 \text{ s}^{-1}$ for the wild-type RGS9 domain and $0.025 \pm 0.002 \text{ s}^{-1}$ for the W299R mutant, based on one of two similar experiments.

s.e.m.), based on two independent experiments; see Fig. 1b and associated legend). These data indicate that the W299R mutation causes a dramatic reduction in the catalytic activity of RGS9, probably by destabilizing its structure.

The fifth index patient (from Guatemala) with an abnormal photoresponse recovery had no detected mutations in *RGS9*. However, evaluation of the *R9AP* gene revealed the homozygous frameshift mutation R65 (1-base-pair insertion) (CGG to CCGG; c.194insC). If translated, this allele would encode a 298-amino-acid protein containing only 65 amino-terminal residues of R9AP. This is an obviously null allele because the carboxy-terminal ~3/4 of the R9AP molecule will be replaced by a random sequence. The mutant is highly unlikely to bind RGS9 or to have a transmembrane anchoring domain. This patient was the offspring of consanguineous parents (the parents shared 3/64 of their genes by descent; see Supplementary Information for pedigree structure). The patient's two siblings were unaffected heterozygotes. Neither this *R9AP* mutation nor the W299R *RGS9* mutation were found among 93 control individuals or among screens of 283 and 614 patients, respectively, with other hereditary retinal diseases.

The chief complaint of all *RGS9* and *R9AP* patients was difficulty adjusting to changes in luminance. Walking out of a house into the sunlight caused temporary blindness so severe that it necessitated immobility or assistance for 5–10 s. Visual compromise was also experienced when travelling from a bright environment to a dark one (for example, daytime driving into a tunnel). Patients tended to avoid environments with high luminance; such environments required frequent blinking or dark sunglasses. Some patients stated that they could not participate in ball games because they could not see a moving ball. The patients' symptoms were present from early childhood. Neither the patients' symptoms nor their ocular findings including ERGs (see below) appreciably changed during repeat evaluations extending up to 14 yr, suggesting that they may have a stationary condition.

Standard visual acuity testing (high-contrast letters) revealed normal to subnormal acuities (20/20 to 20/80), with the measured acuity fluctuating from visit to visit. Optically corrected visual acuity increased in most patients by a factor of two with the use of pinholes, possibly owing to a reduction in retinal illuminance. One of the *RGS9* index patients had a unilateral decrease in acuity

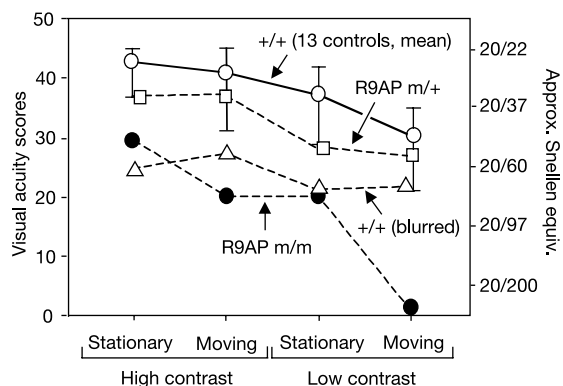


Figure 2 Dynamic acuity. The left axis denotes the total number of correctly identified letters; the approximate Snellen equivalent sizes of the test letters is on the vertical right axis. The results for each individual under the different test conditions are connected with lines. The acuity of the *R9AP* homozygote measured with stationary test letters was slightly less with low-contrast versus high-contrast letters; this contrast-related reduction involving stationary letters was close to that found in an *R9AP* heterozygote and in a group of 13 normal controls (open circles, means; vertical bars, ranges of results). The *R9AP* homozygote has a severe reduction in acuity when low-contrast letters are in motion. This reduction is not seen in a normal control who is blurred with a spectacle lens overcorrection of +1.75 dioptres.

(20/240–20/399) interpreted as amblyopia. The effect of contrast and movement on acuity was evaluated in the *R9AP* patient (Fig. 2). In normal controls, movement of the test letters resulted in only a small decrease in acuity. In the *R9AP* patient, however, movement had a pronounced effect with the most severe reduction in acuity (to a level corresponding to less than 20/200) occurring with moving, low-contrast letters. The relationship of visual acuity to prior light exposure was evaluated in the *R9AP* patient using a photostress test. Following direct ophthalmoscopic presentation of a 9° diameter spot of light centred on the fovea for 10 s through a dilated pupil¹⁸, the patient required 115 s to recover central visual acuity versus 11.2 ± 3.4 s (mean $\pm$ s.d.) in controls; an *R9AP* heterozygote had a normal recovery time (11 s).

Colour vision, tested with Ishihara plates, the Farnsworth D-15 panel (saturated and desaturated), and the Farnsworth–Munsell 100-hue test, was normal (not every test was performed on each patient). The *R9AP* patient had full visual fields tested with the Goldman perimeter V-4e test light and normal sensitivity tested with the Humphrey field analyser 30-2 threshold program. The two *RGS9* patients whose visual fields were evaluated had a generalized slight reduction (0.3 log-units) in mean sensitivity tested with the Humphrey field analyser 30-2 threshold test. Final dark-adaptation thresholds were normal. Ophthalmoscopy showed no abnormalities in the patients' fundi.

Routine ERGs showed normal rod responses to flashes of dim blue light (Fig. 3). There was no attenuation in ERG amplitudes in response to subsequent dim blue flashes separated by 2 s (0.5 Hz). Rod-plus-cone ERGs to 0.5-Hz bright white flashes also showed a normal response, but only to the first flash (Fig. 3). The responses to the second flash and all subsequent flashes were markedly reduced in amplitude (Fig. 3). As the majority of the ERG amplitude in response to this flash intensity is normally mediated by rods^{19,20}, this reduction in amplitude suggests a rod dysfunction. This interpretation is supported by the fact that the responses to the subsequent flashes had short implicit times and waveforms similar to cone responses in patients with congenital stationary night blindness due to defects confined to the rod phototransduction pathway²¹. Thus, it is likely that the responses from the second and subsequent flashes were mediated primarily by the cone mechanism and that rods become more light-adapted than cones at this stimulus intensity. A dysfunction also involving cones in these patients is revealed by the cone ERG responses elicited by 30-Hz white light flashes (Fig. 3). Early responses were of normal amplitude and timing, but the responses rapidly decreased in amplitude so that the 30-Hz ERGs became non-detectable without computer averaging in about 2 s.

To measure the temporal aspects of the recovery of retinal function, we recorded ERG responses to pairs of white flashes with different inter-flash intervals (double-flash ERGs) in both

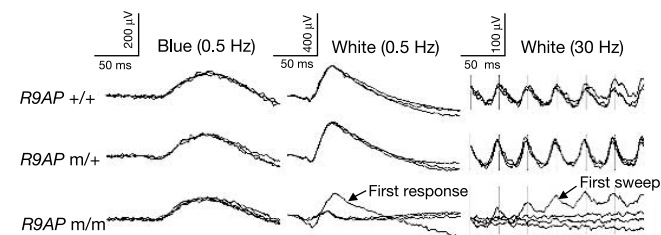


Figure 3 Full-field ERGs. Data are from a normal control (+/+), an *R9AP* heterozygote (m/+), and an *R9AP* homozygote (m/m) in response to 0.5-Hz dim blue flashes to elicit a rod response (left column), to 0.5-Hz white flashes to elicit a near-maximal rod-plus-cone response (middle column), and the same white flash repeated at 30 Hz to isolate a cone response (right column). Vertical lines in the 30-Hz responses designate flash onsets. For each test condition, four consecutive recordings are shown.

R9AP and *RGS9* patients (Fig. 4). In normal controls, the amplitudes of the a-wave and the b-wave of the second flash were normal if the inter-flash interval was 2 s or greater. In the patients, there was a profound suppression of the second-flash ERG, requiring up to 1 min for full recovery (Fig. 4). The suppression of the second flash is similar to that previously reported in *Rgs9* knockout mice¹². A suppression of the second flash ERG was observed in the *R9AP* heterozygote with inter-flash intervals of 1 or 2 s (Fig. 4).

On the basis of patients' symptoms and vision test results, it appears that their photoreceptors require an abnormally long time to adjust to changes in luminance. Slow adaptation to dim light, as exhibited by the abnormal photostress test, is also experienced by some patients who have retinitis pigmentosa or congenital stationary night blindness. What is unusual about the patients described here is that they are also slow to adapt to bright light. The slow

adaptation of the photoreceptors results in a transient, incapacitating blindness when patients are suddenly confronted with a brightly lit environment. The photoreceptor defect also results in a severe reduction in acuity if low-contrast objects move against a bright background. The high conservation of *RGS9* among vertebrates suggests that these visual abnormalities have been strictly selected against during evolution. A previous report of patients now known to be *RGS9* mutant homozygotes labelled their disease according to their abnormal double-flash ERGs (prolonged electro-retinal response suppression, PERRS)¹⁶. With the recognition of the underlying photoreceptor defect and a better understanding of the patients' difficulty in adapting to changes in luminance, we propose the term bradyopsia ('slow vision'; Greek) for this clinical entity. □

Methods

Single turnover GTPase assay

The W299R mutation was introduced using PCR-based methods, and single-turnover transducin GTPase assays were conducted as described²² using 3 μ M *RGS9* domains, both expressed as an N-terminal GST fusion proteins²³. The γ -subunit of cGMP phosphodiesterase (2 μ M) was added in the assays to maximize the *RGS9* catalytic activity.

Dynamic acuity

From a distance of 0.9 m, each patient or control (wearing an optimal spectacle correction if necessary) was presented with black letters of high contrast (90%) or low contrast (10%) against a white background on a computer monitor (luminance 130 cd m^{-2}). Stationary letters appeared in the middle of the screen; moving letters travelled horizontally across the monitor (one sweep, left to right) at $18.5^\circ \text{ s}^{-1}$. Letters were presented one at a time. The first and largest set of five letters corresponded approximately to 20/200 Snellen acuity at the test distance; each subsequent set of five letters was approximately 21% smaller than the previous set; letters in the last (ninth) set were equivalent to 20/22. The test concluded when one entire set of five letters was missed.

Electrophysiology

Full-field ERGs were performed after maximally dilating pupils with eyedrops (1% cyclopentolate, 0.25% tropicamide and 2.5% phenylephrine), and dark-adaptation for at least 40 min. Waveforms were recorded with a Burian–Allen contact-lens electrode as previously reported^{24,25}. For the 30-Hz ERG, the first sweep was recorded starting approximately 1 s after the start of the light flashes, and each successive sweep was separated from the last sweep by approximately 1 s. The homozygote's responses to the first sweep of 30-Hz white light flashes were larger than the responses to subsequent sweeps. Double-flash ERGs were begun after 45 min of dark adaptation, and were elicited with 10- μ s full-field flashes of white light (1.6 cd s m^{-2} ; equivalent to the ISCEV standard flash)^{26,27}. Between each pair of flashes there was a dark-adaptation interval of 2 min. High-frequency oscillations following the cornea-positive (b-wave) peak in some tracings represent blink artefacts.

Received 27 July; accepted 14 October 2003; doi:10.1038/nature02170.

- Dohlman, H. G., Song, J., Ma, D., Courchesne, W. E. & Thorne, J. Sst2, a negative regulator of pheromone signaling in the yeast *Saccharomyces cerevisiae*: Expression, localization, and genetic interaction and physical association with Gpa1 (the G-protein α subunit). *Mol. Cell. Biol.* **16**, 5194–5209 (1996).
- Rogers, J. H. et al. *RGS4* causes increased mortality and reduced cardiac hypertrophy in response to pressure overload. *J. Clin. Invest.* **104**, 567–576 (1999).
- Oliveira-dos-Santos, A. J. et al. Regulation of T cell activation, anxiety, and male aggression by *RGS2*. *Proc. Natl Acad. Sci. USA* **97**, 12272–12277 (2000).
- Chen, C. K. et al. Slowed recovery of rod photoreceptor in mice lacking the GTPase accelerating protein *RGS9-1*. *Nature* **403**, 557–560 (2000).
- Sinnarajah, S. et al. *RGS2* regulates signal transduction in olfactory neurons by attenuating activation of adenylyl cyclase III. *Nature* **409**, 1051–1055 (2001).
- Heximer, S. P. et al. Hypertension and prolonged vasoconstrictor signaling in *RGS2*-deficient mice. *J. Clin. Invest.* **111**, 445–452 (2003).
- He, W., Cowan, C. W. & Wensel, T. G. *RGS9*, a GTPase accelerator for phototransduction. *Neuron* **20**, 95–102 (1998).
- Cowan, C. W., Fariss, R. N., Sokal, I., Palczewski, K. & Wensel, T. G. High expression levels in cones of *RGS9*, the predominant GTPase accelerating protein of rods. *Proc. Natl Acad. Sci. USA* **95**, 5351–5356 (1998).
- Hu, G. & Wensel, T. G. *R9AP*, a membrane anchor for the photoreceptor GTPase accelerating protein, *RGS9-1*. *Proc. Natl Acad. Sci. USA* **99**, 9755–9760 (2002).
- Lishko, P. V., Martemyanov, K. A., Hopp, J. A. & Arshavsky, V. Y. Specific binding of *RGS9-G β 5L* to protein anchor in photoreceptor membranes greatly enhances its catalytic activity. *J. Biol. Chem.* **277**, 24376–24381 (2002).
- Hu, G., Zhang, Z. & Wensel, T. G. Activation of *RGS9-1* GTPase acceleration by its membrane anchor, *R9AP*. *J. Biol. Chem.* **278**, 14550–14554 (2003).
- Lyubarsky, A. L. et al. *RGS9-1* is required for normal inactivation of mouse cone phototransduction. *Mol. Vis.* **7**, 71–78 (2001).
- Martemyanov, K. A. et al. The DEP domain determines subcellular targeting of the GTPase activating protein *RGS9* in vivo. *J. Neurosci.* **23**, 10175–10181 (2003).

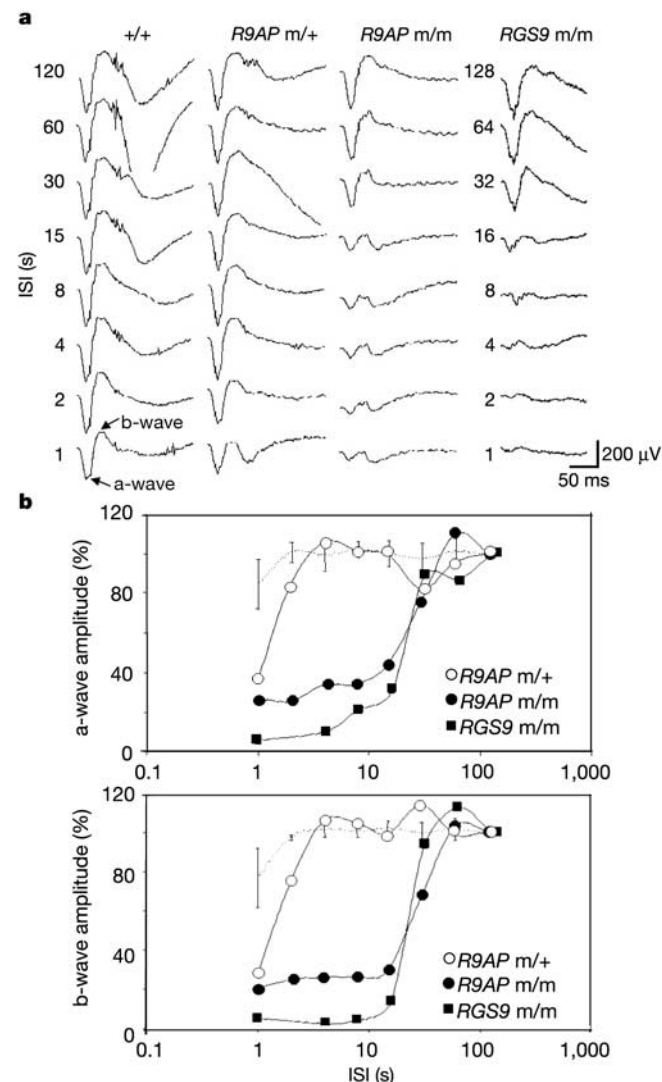


Figure 4 Double-flash ERGs. **a**, The responses to the second flash of a pair of flashes with different interstimulus intervals (ISIs). The *RGS9* homozygote was patient 2 in a previous report¹⁶. The ISIs at left are for the first three columns; the ISIs next to the right column are for that column. **b**, Plots of the amplitudes of the a-wave (top) and b-wave (bottom) of the responses to the second flash of a pair of flashes, expressed as a percentage of the mean baseline amplitude for the respective patient or control individual; that is, for ISI = 120 s. Data are the means of both eyes in the *R9AP* and *RGS9* homozygotes and from a single eye of the *R9AP* heterozygote. Error bars connected by a dotted line show the mean $\pm$ s.d. of the results from one eye of normal controls ($n = 6$ for the a-wave, and $n = 5$ for the b-wave).

14. Zhang, K. *et al.* Structure, alternative splicing, and expression of the human RGS9 gene. *Gene* **240**, 23–34 (1999).
15. Rahman, Z. *et al.* Cloning and characterization of RGS9-2: A striatal-enriched alternatively spliced product of the RGS9 gene. *J. Neurosci.* **19**, 2016–2026 (1999).
16. Kooijman, A. C., Houtman, A., Damhof, A. & van Engelen, J. P. Prolonged electro-retinal response suppression (PERRS) in patients with stationary subnormal visual acuity and photophobia. *Doc. Ophthalmol.* **78**, 245–254 (1991).
17. Slep, K. C. *et al.* Structural determinants for regulation of phosphodiesterase by a G protein at 2.0 Å. *Nature* **409**, 1071–1077 (2001).
18. Sandberg, M. A., Pawlyk, B. S. & Berson, E. L. Acuity recovery and cone pigment regeneration after a bleach in patients with retinitis pigmentosa and rhodopsin mutations. *Invest. Ophthalmol. Vis. Sci.* **40**, 2457–2461 (1999).
19. Berson, E. L., Gouras, P. & Gunkel, R. D. Progressive cone degeneration, dominantly inherited. *Arch. Ophthalmol.* **80**, 77–83 (1968).
20. Kooijman, A. C. *et al.* Spectral sensitivity function of the a-wave of the human electroretinogram in the dark-adapted eye. *Clin. Vision Sci.* **6**, 379–383 (1991).
21. Sandberg, M. A. *et al.* Rod and cone function in the Nougaret form of stationary night blindness. *Arch. Ophthalmol.* **116**, 867–872 (1998).
22. Martemyanov, K. A. & Arshavsky, V. Y. Noncatalytic domains of RGS9-1-Gβ5L play a decisive role in establishing its substrate specificity. *J. Biol. Chem.* **277**, 32843–32848 (2002).
23. He, W. *et al.* Modules in the photoreceptor RGS9-1-Gβ5L GTPase-accelerating protein complex control effector coupling, GTPase acceleration, protein folding, and stability. *J. Biol. Chem.* **275**, 37093–37100 (2000).
24. Berson, E. L., Gouras, P. & Gunkel, R. D. Rod responses in retinitis pigmentosa, dominantly inherited. *Arch. Ophthalmol.* **80**, 58–66 (1968).
25. Andreasson, S. O., Sandberg, M. A. & Berson, E. L. Narrow-band filtering for monitoring low-amplitude cone electroretinograms in retinitis pigmentosa. *Am. J. Ophthalmol.* **105**, 500–503 (1988).
26. Marmor, M. F. An international standard for electroretinography. *Doc. Ophthalmol.* **73**, 299–302 (1989).
27. Marmor, M. F. & Zrenner, E. Standard for clinical electroretinography (1994 update). *Doc. Ophthalmol.* **89**, 199–210 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Li, T. McGee and B. Pawlyk for assistance. This study followed the tenets of the Declaration of Helsinki; it was approved by the Human Studies Committees of the authors' institutions and all patients gave their consent before their participation. This work was supported by the NIH, the Foundation Fighting Blindness, Owings Mills, Maryland, and the American Heart Association.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.P.D. (Thaddeus_Dryja@meei.harvard.edu).

Expression and function of orphan nuclear receptor TLX in adult neural stem cells

Yanhong Shi¹, D. Chichung Lie², Philippe Taupin², Kinichi Nakashima^{2,4}, Jasodhara Ray², Ruth T. Yu¹, Fred H. Gage² & Ronald M. Evans^{1,3}

¹Gene Expression Laboratory, ²Laboratory of Genetics, and ³Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

⁴Department of Cell Fate Modulation, Institute of Molecular Embryology and Genetics, Kumamoto University, 2-2-1, Honjo, Kumamoto 860-0811, Japan

The finding of neurogenesis in the adult brain led to the discovery of adult neural stem cells¹. TLX was initially identified as an orphan nuclear receptor expressed in vertebrate forebrains² and is highly expressed in the adult brain³. The brains of TLX-null mice have been reported to have no obvious defects during embryogenesis⁴; however, mature mice suffer from retinopathies⁵, severe limbic defects, aggressiveness, reduced copulation and progressively violent behaviour^{4,6}. Here we show that TLX maintains adult neural stem cells in an undifferentiated, proliferative state. We show that TLX-expressing cells

isolated by fluorescence-activated cell sorting (FACS) from adult brains can proliferate, self-renew and differentiate into all neural cell types *in vitro*. By contrast, TLX-null cells isolated from adult mutant brains fail to proliferate. Reintroducing TLX into FACS-sorted TLX-null cells rescues their ability to proliferate and to self-renew. *In vivo*, TLX mutant mice show a loss of cell proliferation and reduced labelling of nestin in neurogenic areas in the adult brain. TLX can silence glia-specific expression of the astrocyte marker GFAP in neural stem cells, suggesting that transcriptional repression may be crucial in maintaining the undifferentiated state of these cells.

Expression of TLX in the mouse starts at embryonic day 8 (E8), peaks at E13.5 and decreases by E16, with barely detectable levels at birth. TLX expression increases after birth and is high in the adult brain³. Although TLX-null mice appear grossly normal at birth, mature mice manifest a rapid retinopathy⁵ with reduced cerebral hemispheres^{4,6}. Histologically, adult mutant brains have severely reduced hippocampal dentate gyri, greatly expanded lateral ventricles and reduced olfactory bulbs (Fig. 1a), all of which are active adult neurogenic areas⁷. Behaviourally, TLX mutants show aggressiveness, reduced copulation and progressively violent behaviour^{4,6}. The hypomorphic defects and neurological disorders of the mutant mice suggest that TLX has a role in normal function of the central nervous system (CNS).

Taking advantage of a β-galactosidase (β-gal) reporter, which was knocked into the TLX locus, we examined the expression pattern of TLX in adult brains of heterozygote mice. LacZ staining was distributed sparsely throughout the cortex, but indicated high but dispersed expression of TLX in the subgranular layer of the dentate gyri and clustered expression in the subventricular zone (SVZ; Fig. 1b). Notably, these are the two main sites where adult neural stem cells (NSCs) are located⁷. Previous studies suggested that nestin is a common marker of proliferating CNS progenitors^{8,9}. Immunohistological assessment of brain sections from adult TLX^{+/-} mice showed the colocalization of β-gal and nestin in both the dentate gyri and the SVZ (Fig. 1c), suggesting that TLX is expressed in adult NSCs or progenitor cells.

An emerging concept contends that a subset of the stem cell pool corresponds to a repository of relatively quiescent cells that serve as the source for actively dividing cells¹⁰. To examine whether TLX-expressing cells represent the quiescent or dividing cells, we carried out β-gal and 5-bromodeoxyuridine (BrdU) double staining on brain sections from TLX^{+/-} mice treated with BrdU. The analysis showed that TLX expression corresponds to both BrdU-positive and BrdU-negative cells in the adult germinal zones (dentate gyri and SVZ; Fig. 1d).

Cell sorting was used to isolate TLX-expressing cells to determine their ability to proliferate, self-renew and give rise to both neurons and glia. By using a fluorogenic LacZ substrate, β-gal-positive cells were isolated from TLX^{+/-} forebrains and cultured in N2 medium supplemented with epidermal growth factor (EGF), fibroblast growth factor (FGF) and heparin¹¹. Immunostaining confirmed expression of β-gal in the FACS-sorted cells. The proliferation potential of TLX-positive cells was examined by BrdU labelling of dividing cells. BrdU treatment for 24 h and subsequent immunostaining showed that more than 98% of the β-gal-positive cells were BrdU- and nestin-positive (Fig. 2a).

We next tested the self-renewal capacity of TLX-expressing cells by using clonal analysis¹². Clonal populations were derived from single FACS-sorted, β-gal-positive cells. From a total of 28 single cells, 12 clones developed, 8 of which reached more than 200 cells by day 12; another 4 reached this number by day 25. The clones were dissociated and single cells were plated for a second round of cloning. All of the clones tested were also capable of secondary expansion. A prototypical clonal expansion is shown in Fig. 2b. These results show that TLX-expressing cells comprise a self-renewing population.

We then tested whether TLX cells derived from either primary or secondary clones could be induced to differentiate. Indeed, when examined for the expression of Tuj1 (a neuronal marker), GFAP (an astrocyte marker) and O4 (an oligodendrocyte marker), all three neural cell types with characteristic morphologies were generated upon differentiation (Fig. 2c), indicating that TLX-expressing cells are multipotent. By contrast, the cells isolated from the forebrains of TLX-null littermates failed to proliferate in the same growth conditions (Fig. 2d and Supplementary Fig. 1). Immunostaining showed that, whereas the TLX^{+/-} cells are nestin-positive and GFAP-negative, the TLX^{-/-} cells are nestin-negative but GFAP-

positive, suggesting the spontaneous differentiation of astrocytes (Fig. 2d and Supplementary Fig. 2).

In an attempt to rescue the proliferative defect, FACS-sorted adult TLX^{-/-} (LacZ/LacZ) cells were infected with a lentiviral vector expressing both TLX and green fluorescent protein (GFP). The resulting infected cells were re-sorted on the basis of the internal GFP marker. These infected and selected cells regained the ability to proliferate and could be clonally expanded (Fig. 2e, f). Expression of TLX led to a substantial restoration of cell proliferation, as detected by staining for nestin (Supplementary Fig. 3) and Ki67 (a proliferative marker), and reduced astrocyte differentiation, as detected

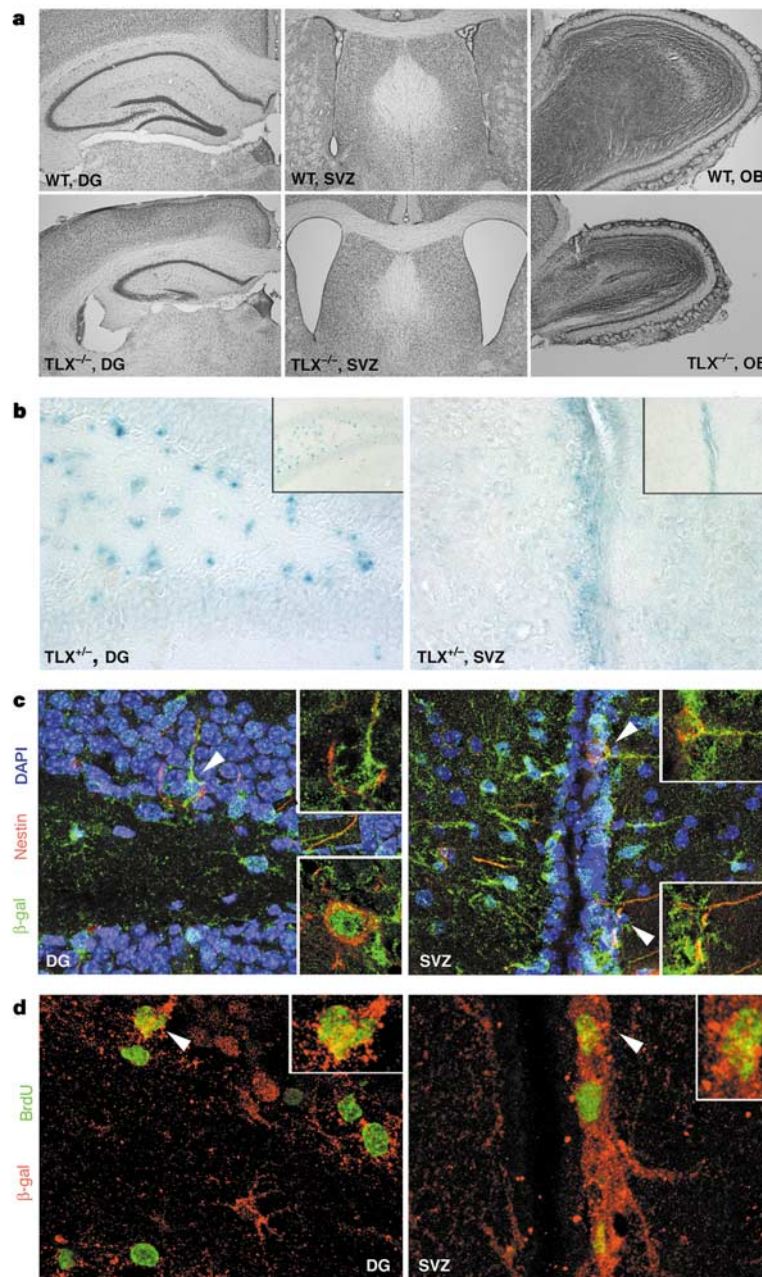


Figure 1 Expression of TLX in adult NSCs. **a**, Nissl staining of brain sections from adult wild-type (WT) and mutant mice. **b**, LacZ staining of dentate gyri and SVZ sections from adult TLX^{+/-} mice. Insets show images at lower magnification. **c**, Co-staining of β -gal and nestin in dentate gyri and SVZ sections from TLX^{+/-} mice. Insets show enlargements of the double-stained cells indicated by arrowheads. The inset in the lower right of the

dentate gyri section shows an example of nestin staining around the nuclei. **d**, Co-staining of β -gal and BrdU in dentate gyri and SVZ sections from TLX^{+/-} mice. Insets show enlargements of the double-stained cells indicated by arrowheads. DG, dentate gyri; OB, olfactory bulbs.

by staining for GFAP (Fig. 2e and Supplementary Fig. 4). By contrast, cells that were infected with a GFP control virus underwent spontaneous differentiation (data not shown). In addition, clonal analysis showed that the cells expressing viral TLX could be clonally expanded from single GFP-positive cells and were both nestin and GFP positive (Fig. 2f). Because the lentivirus-expressed TLX is controlled by the constitutive cytomegalovirus (CMV) promoter, the rescued cells continued to proliferate even under differentiation conditions, as expected (data not shown). Together, these results show that TLX can rescue the undifferentiated, proliferative state of NSCs *in vitro*.

We next examined the transcriptional property of TLX to address how it might contribute to NSC maintenance. When fused to the DNA-binding domain (DBD) of GAL4, TLX strongly repressed a luciferase reporter that was downstream of GAL4 DNA-binding

sites (data not shown). In searching for downstream targets of TLX, a substantial upregulation of GFAP, S100 β and aquaporin 4 (AQP4) was detected in the TLX mutant brains. TLX was expressed in the proliferating NSCs but was switched off upon differentiation, whereas GFAP, S100 β and AQP4 were expressed after differentiation (Fig. 3a). The case for direct regulation was strengthened by the identification of a consensus TLX-binding site, AAGTCA, in the promoters of these genes. Gel shift analysis showed that TLX could specifically bind to these sequences *in vitro* (Fig. 3b and data not shown), suggesting that astrocyte-specific genes are direct downstream targets of TLX repression.

To confirm further this repression by TLX, we carried out reporter assays using a GFAP-promoter-driven luciferase reporter (GFAP-luc)¹³. Leukaemia inhibitory factor (LIF) has been shown to induce astrocyte differentiation and GFAP expression¹³. Co-

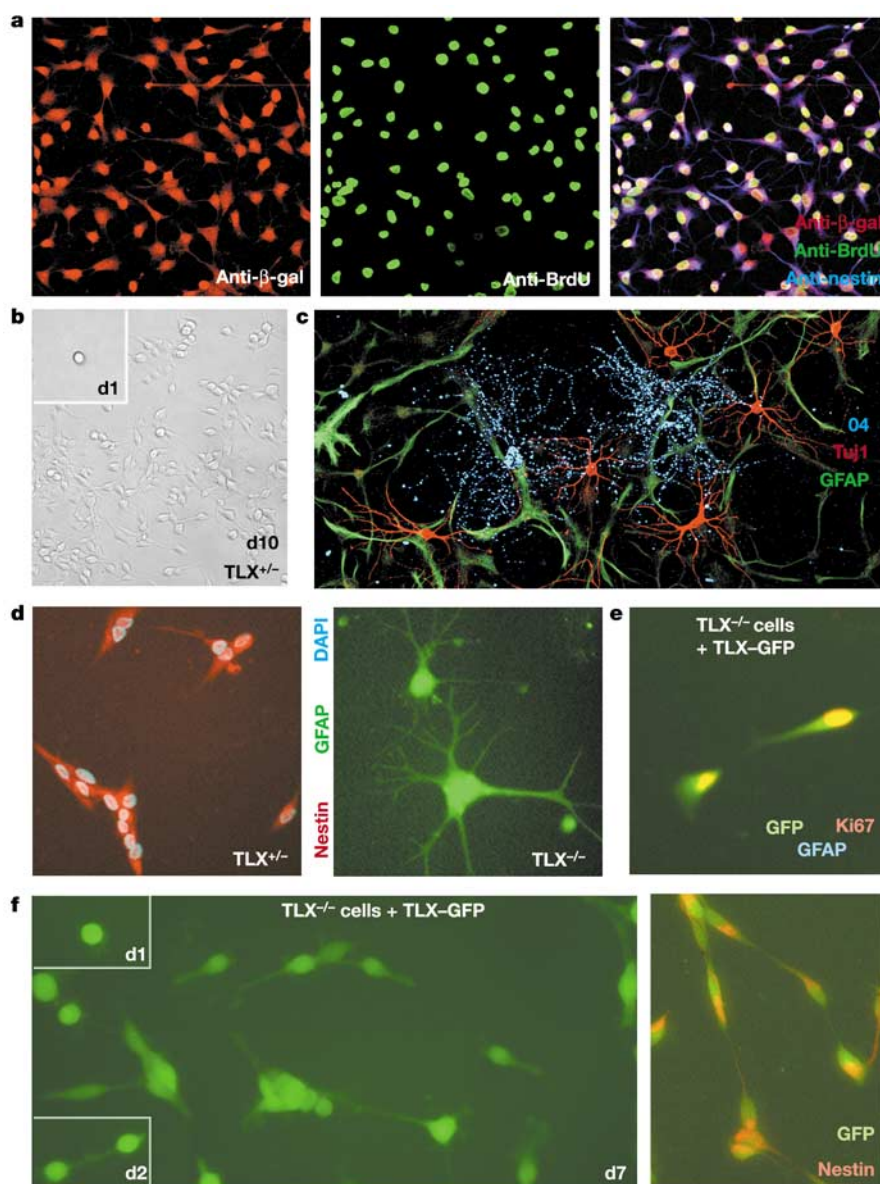


Figure 2 TLX-positive cells are self-renewable and multipotent. **a**, TLX-positive cells were sorted by FACS, treated with BrdU and immunostained for β -gal, BrdU and nestin. **b**, A clonal population of TLX-positive cells at days 1 and 10. **c**, Clonally derived cells were differentiated and stained for Tuj1, GFAP and O4. **d**, Immunostaining of TLX^{+/−} and TLX^{−/−} cells cultured in growth media (see Supplementary Fig. 2 for quantification).

e, Immunostaining of TLX^{−/−} cells expressing TLX-GFP. GFAP staining is negative (see Supplementary Fig. 4 for quantification). **f**, Left, clonal analysis of TLX^{−/−} cells expressing TLX-GFP at days 1, 2 and 7. Right, GFP fluorescence and nestin staining in the cloned cells.

transfection of TLX led to a considerable repression (4.8-fold) of LIF-induced GFAP reporter activity (Fig. 3c), similar to the repression mediated by a dominant-negative variant of STAT3 (ref. 13), which was used as a positive control.

To establish further the role of TLX in the repression of GFAP expression and astrocyte differentiation, we infected NSCs with the TLX-expressing retrovirus. Because it is under a constitutive CMV promoter, the viral TLX remains expressed upon differentiation, in contrast to endogenous TLX, which is downregulated in differentiated cells. As expected, the expression of GFAP and two other astrocyte-specific markers (S100 β and AQP4) was substantially decreased in the cells expressing viral TLX (Fig. 3d). Differentiation of these cells into GFAP-positive astrocytes was reduced proportionately (Fig. 3e). Immunostaining (Fig. 3f) further showed that NSCs with viral TLX expression failed to differentiate into GFAP-positive astrocytes upon treatment with LIF and bone morphogenetic protein 2 (BMP2), a condition favouring astrocyte differentiation. Instead, they continued expressing the neural progenitor marker nestin. Control NSCs infected with GFP virus differentiated into GFAP-positive astrocytes and lost nestin expression under the same conditions. These results further strengthen the argument that TLX is involved in the repression of astrocyte differentiation.

The above analyses suggest that TLX has a role in maintaining the undifferentiated, proliferative state of NSCs. Immunohistochemical assessment of the adult germinal zones detected a marked reduction of nestin-positive cells in both the hippocampal dentate gyri and the SVZ of adult TLX-null mice (Fig. 4a–c), indicating reduced neural precursors in the mutant neurogenic areas. Despite the hypotrophic nature of the mutant brains, an increase in GFAP staining was observed, consistent with the northern analysis (Fig. 3a) and suggesting that there is increased glia differentiation in the brains of mutant mice. TdT-mediated dUTP nick end labelling (TUNEL) assays detected no significant differences in cell death between wild-type and mutant brains (data not shown), suggesting a failure of ongoing cell proliferation or a very early window of cell death.

Next, we examined whether TLX affects cell proliferation in the adult brain. BrdU labelling of adult mice was carried out for 1 week, followed by immunohistochemical assessment. Intensive BrdU labelling was observed along the subgranular zone of the dentate gyri and in the SVZ of the TLX^{+/+} brains, whereas virtually no BrdU labelling was detected in the mutant hippocampus and SVZ (Fig. 4d, e). Quantification of the BrdU-positive cells in the dentate gyri is shown in Fig. 4f. By contrast, as shown by S100 β staining, increased numbers of glia cells were observed in the mutant brains

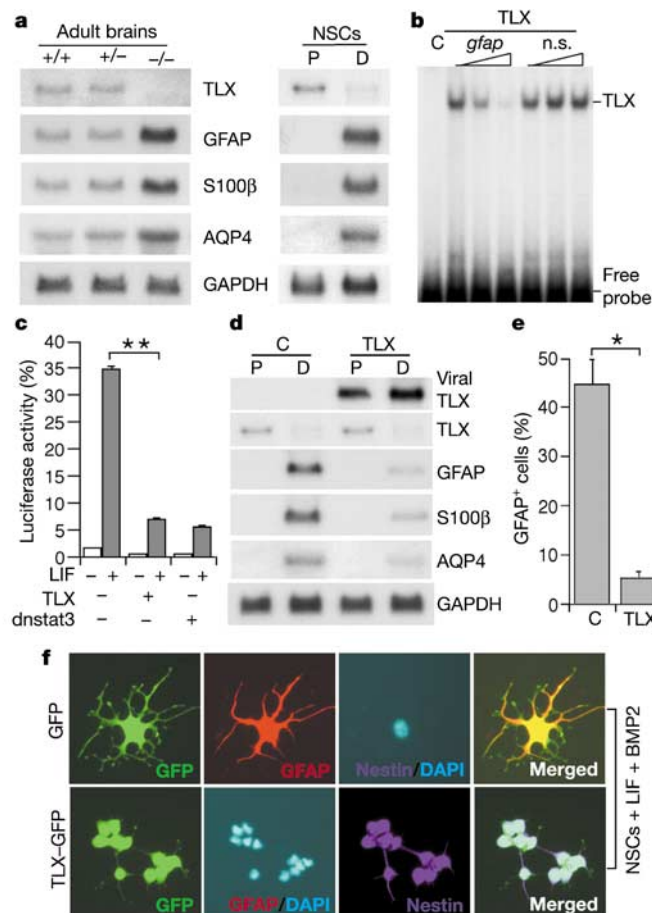


Figure 3 TLX is a transcriptional repressor. **a**, Gene expression in the forebrain of TLX^{+/+}, TLX^{+/-} and TLX^{-/-} mice and in NSCs cultured in proliferation (P) or differentiation (D) conditions analysed by northern blotting. **b**, TLX binds to the GFAP promoter in a gel-shift assay. TLX binding is competed by unlabelled *gfap* but not nonspecific (n.s.) oligos. C, control lysates. **c**, TLX represses GFAP in reporter assays. LIF-induced GFAP-luciferase activity was measured in the absence or the presence of TLX expression. ***P* < 0.0001

by Student's *t*-test. **d**, Gene expression in control (C) or viral TLX-expressing NSCs analysed by northern blotting. P, proliferation; D, differentiation. **e**, Quantification of GFAP-positive astrocytes in control or viral TLX-expressing cells upon differentiation. **P* < 0.001 by Student's *t*-test. **f**, Immunostaining of rat NSCs expressing viral GFP or TLX-GFP.

(Fig. 4g). These results further support our proposal that TLX is essential in maintaining the undifferentiated, proliferative state of stem cells in adult brains.

What are the molecular determinants of adult NSCs? Our characterization of TLX suggests that it may be one of the key regulators that act by controlling the expression of a network of target genes to establish the undifferentiated and self-renewable state of NSCs. The expression of TLX in scattered cells throughout the cortex suggests that it may have another role in addition to its function in the germinal zone. Furthermore, the dominant role in stem cell maintenance that TLX has in the adult is clearly different from its seemingly more modest role in development, although this does not exclude the possibility that the lack of adult NSCs may also

be influenced by developmental functions of TLX. With regard to the persistence of NSCs in the absence of TLX in the adult, the fact that the adult rescue experiment works at all indicates either that a dormant progenitor or NSC population exists or that TLX expression alone can recover a stem cell state in a cell population obtained from the adult germinal region. The characterization of the TLX-expressing cells provides a means to elucidate the molecular and cellular mechanisms underlying stem cell proliferation and differentiation. Furthermore, the use of β -gal-based FACS facilitates the isolation of NSCs from adult brains, which will make it possible to explore clinical applications for transplantability and molecular engineering associated with the treatment of neurodegenerative diseases and brain injuries. □

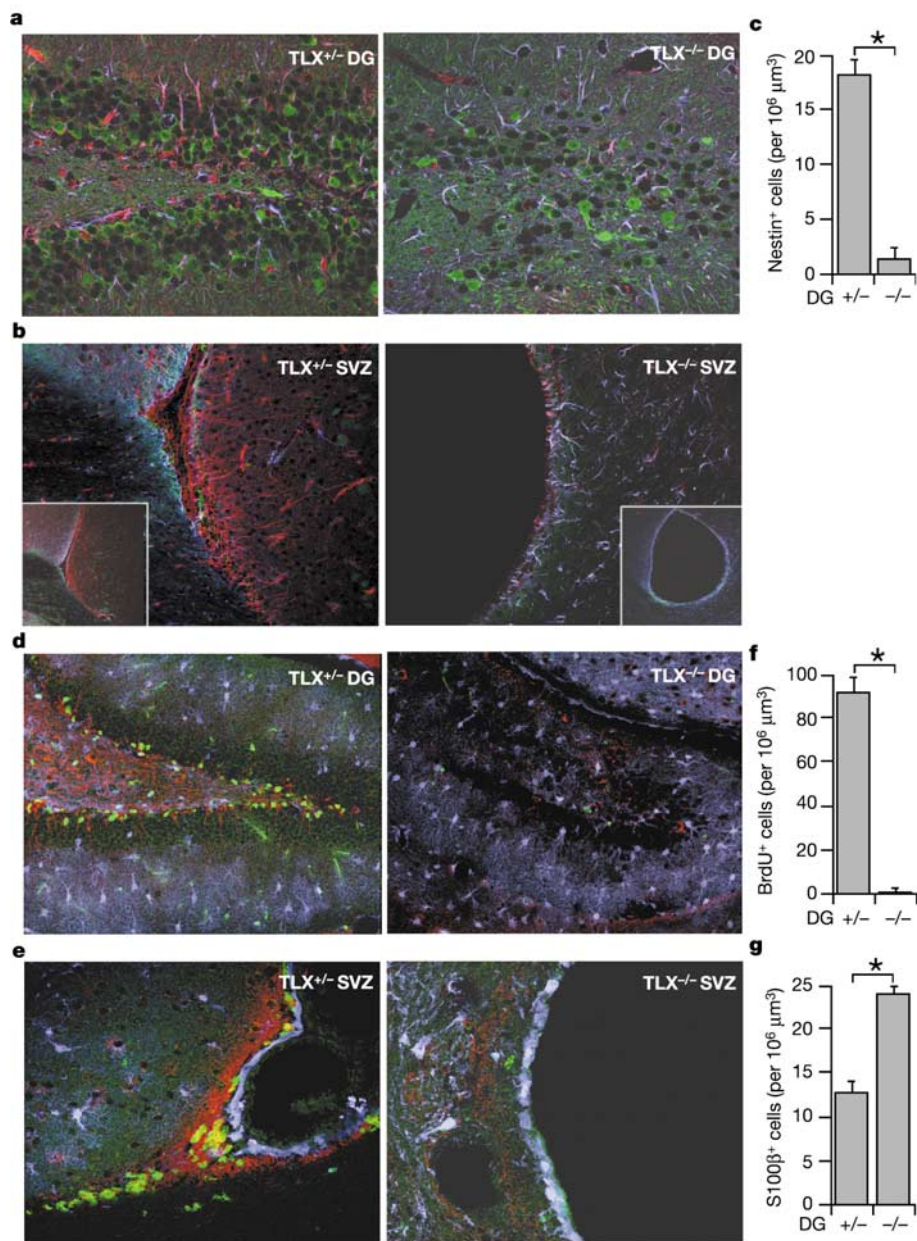


Figure 4 Loss of nestin and BrdU staining in TLX mutant mouse brain. **a**, Coronal dentate gyri sections from the brains of TLX^{+/+} and TLX^{-/-} mice were immunostained for nestin (red), calbindin (green) and GFAP (blue). **b**, Coronal SVZ sections stained for nestin (red), Tuj1 (green) and GFAP (blue). **c**, Quantification of nestin-positive cells in dentate gyri of TLX^{+/+} and TLX^{-/-} mice. **P* < 0.01 by Student's *t*-test. **d**, **e**, Hippocampal dentate gyri

(**d**) or SVZ (**e**) sections from BrdU-treated TLX^{+/+} and TLX^{-/-} mice were immunostained for BrdU (green), Tuj1 (red) and S100β (blue). **f**, **g**, Quantification of BrdU-positive cells (**f**) or S100β-positive cells (**g**) in dentate gyri of TLX^{+/+} and TLX^{-/-} mice. **P* < 0.001 by Student's *t*-test.

Methods

Immunocytochemistry and quantification

Immunostaining was carried out on 40- μ m free-floating sections¹⁴ or with cultured cells¹² as described. We used the following primary antibodies: rat anti-BrdU (Accurate; diluted 1:250), mouse anti-nestin (Pharmingen; 1:1,000), mouse anti-Tuj1 (Bavco; 1:1,000), rabbit anti-S100 β (Sigma; 1:500), guinea-pig anti-GFAP (Advance Immuno; 1:500), mouse O4 IgM (a gift from O. Boegler; 1:3), rabbit anti- β -gal (Cortex; 1:2,000), rabbit anti-Ki67 (Novocastra; 1:1,000). BrdU-labelled cells were treated in 1 M HCl at 37 °C for 30 min and visualized by a confocal microscope (BioRad). Quantitative studies were based on four or more replicates. The dentate gyri subgranular areas were traced by semi-automatic stereology using StereoInvestigator software (MicroBrightfield), and volumes were determined by the areas multiplied by the thickness of the sections. Cell densities were calculated by dividing cell numbers by the volume. Statistical analysis was done with Microsoft Excel. For LacZ staining, the 40- μ m frozen sections were fixed with glutaraldehyde and paraformaldehyde and stained in 5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside (X-gal; Gibco-BRL) solution containing 20 mM K-ferricyanide, 20 mM ferrocyanide, 0.01% sodium deoxycholate, 0.02% Nonidet P-40 and 2 mM MgCl₂. We used 40X or 10X objectives for the images. Nissl staining was done by using 0.25% Cresyl Violet on 40- μ m coronal sections.

Preparation and culture of adult NSC or progenitor cells

Mouse forebrains were minced with scalpels and digested in PPD (Papain, Dispase II and DNase I) solution. Cells were isolated using Percoll gradients¹⁵ and cultured in DMEM F12 medium with 1 mM L-glutamine, N2 supplement (Gibco-BRL), EGF (20 ng ml⁻¹), FGF2 (20 ng ml⁻¹) and heparin (5 μ g ml⁻¹). Cells were collected by β -gal-based sorting using fluorescein digalactoside as substrate (Molecular Probes). For differentiation, cells were exposed to N2 supplemented medium containing 5 μ M forskolin and 0.5% fetal bovine serum (FBS) for 1 week. Rat NSCs were cultured in N2 medium containing FGF2 (20 ng ml⁻¹) as described¹⁵. Neural differentiation was initiated with N2 supplemented medium containing 1 μ M retinoic acid and 0.5% FBS for 4 d. For astrocyte differentiation, cells were cultured in N2 supplemented medium containing 50 ng ml⁻¹ LIF, 50 ng ml⁻¹ BMP2 and 1% FBS. Clonal analysis was done in conditional CCG medium¹². β -gal-positive cells were plated in 96-well plates. Cells were counted 4 h after plating. We monitored wells containing single cells continuously until they reached more than 200 cells.

In vivo BrdU labelling

Eight-week-old mice were injected intraperitoneally once daily with BrdU at 50 mg per kg (body weight) over a 12-d period. Brains were fixed 1 d after the last injection and processed for immunostaining as described¹⁴. Alternatively, BrdU (1 mg ml⁻¹) was administered to mice in their drinking water for 2 weeks, and brain tissues were processed for immunostaining.

Northern blot analysis and gel shift assay

Total RNA from cultured cells or brain tissues was isolated using Trizol reagent (Life Technologies). We carried out northern blot analyses as described¹⁶. Gel shift assay was done as described² using *in vitro* translated TLX and ³²P-labelled GFAP probes.

Transient transfection assays and immunoprecipitation analysis

CV-1 cells were transiently transfected as described¹⁶. Adult rat neural progenitor cells were transfected by using Transit-LT (Mirus). Transfected neural progenitors were treated with solvent or LIF (80 ng ml⁻¹) for luciferase assays. We normalized the luciferase activity of each sample to the β -gal activity in CV-1 cells and the Renilla luciferase activity in progenitor cells. Transfections were done in triplicate at least three times.

Viral production and infection

A TLX-expressing retrovirus was produced by using either an NIT vector and 293gp cells¹⁵ or the pMY vector¹⁷ and Plat-E cells¹⁸. The TLX-expressing lentivirus was produced using pCSC vector¹⁹ and 293T cells. The transgene in the NIT vector is expressed from a minimal CMV promoter containing six tetracycline operators. The transgene in the pMY or pCSC vector is expressed from a CMV promoter and upstream of an IRES GFP marker. The NSCs were infected by incubating with the virus and 2 μ g ml⁻¹ polybrene (Sigma). We

selected cells expressing NIT-TLX with 400 μ g ml⁻¹ G418. Cells expressing pMY-TLX and pCSC-TLX were sorted by the IRES GFP marker.

Received 28 July; accepted 9 October 2003; doi:10.1038/nature02211.

1. Gage, F. H. Mammalian neural stem cells. *Science* **287**, 1433–1438 (2000).
2. Yu, R. T., McKeown, M., Evans, R. M. & Umeson, K. Relationship between *Drosophila* gap gene tailless and a vertebrate nuclear receptor Tlx. *Nature* **370**, 375–379 (1994).
3. Monaghan, A. P., Grau, E., Bock, D. & Schutz, G. The mouse homolog of the orphan nuclear receptor tailless is expressed in the developing forebrain. *Development* **121**, 839–853 (1995).
4. Monaghan, A. P. *et al.* Defective limbic system in mice lacking the tailless gene. *Nature* **390**, 515–517 (1997).
5. Yu, R. T. *et al.* The orphan nuclear receptor Tlx regulates Pax2 and is essential for vision. *Proc. Natl Acad. Sci. USA* **97**, 2621–2625 (2000).
6. Chiang, M. Y. & Evans, R. M. *Reverse Genetic Analysis of Nuclear Receptors, RXR γ , RAR β , and TLX in Mice* (Dissertation, Univ. California, San Diego, 1997).
7. Gage, F. H. Neurogenesis in the adult brain. *J. Neurosci.* **22**, 612–613 (2002).
8. Lendahl, U., Zimmerman, L. B. & McKay, R. D. CNS stem cells express a new class of intermediate filament protein. *Cell* **60**, 585–595 (1990).
9. Reynolds, B. A., Tetzlaff, W. & Weiss, S. A multipotent EGF-responsive striatal embryonic progenitor cell produces neurons and astrocytes. *J. Neurosci.* **12**, 4565–4574 (1992).
10. Doetsch, F., Petreanu, L., Caille, I., Garcia-Verdugo, J. M. & Alvarez-Buylla, A. EGF converts transit-amplifying neurogenic precursors in the adult brain into multipotent stem cells. *Neuron* **36**, 1021–1034 (2002).
11. Allen, D. M. *et al.* Ataxia telangiectasia mutated is essential during adult neurogenesis. *Genes Dev.* **15**, 554–566 (2001).
12. Taupin, P. *et al.* FGF-2-responsive neural stem cell proliferation requires CCG, a novel autocrine/paracrine cofactor. *Neuron* **28**, 385–397 (2000).
13. Nakashima, K. *et al.* Synergistic signaling in fetal brain by STAT3–Smad1 complex bridged by p300. *Science* **284**, 479–482 (1999).
14. Lie, D. C. *et al.* The adult substantia nigra contains progenitor cells with neurogenic potential. *J. Neurosci.* **22**, 6639–6649 (2002).
15. Palmer, T. D., Markakis, E. A., Willhoite, A. R., Safar, F. & Gage, F. H. Fibroblast growth factor-2 activates a latent neurogenic program in neural stem cells from diverse regions of the adult CNS. *J. Neurosci.* **19**, 8487–8497 (1999).
16. Shi, Y. *et al.* Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. *Genes Dev.* **15**, 1140–1151 (2001).
17. Misawa, K. *et al.* A method to identify cDNAs based on localization of green fluorescent protein fusion products. *Proc. Natl Acad. Sci. USA* **97**, 3062–3066 (2000).
18. Morita, S., Kojima, T. & Kitamura, T. Plat-E: an efficient and stable system for transient packaging of retroviruses. *Gene Ther.* **7**, 1063–1066 (2000).
19. Miyoshi, H., Blomer, U., Takahashi, M., Gage, F. H. & Verma, I. M. Development of a self-inactivating lentivirus vector. *J. Virol.* **72**, 8150–8157 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. Cabrera, S. Tiep, M. Nelson, H. Juguilon, J. Havstad, B. Miller, R. Summers, A. Consiglio, A. Huynh, L. Moore, A. Dearie and H. Lansford for technical help; M. L. Gage for editing; E. Stevens and E. Ong for administrative assistance; C. Zhang for comments on the manuscript; R. Marr and I. Verma for the lentiviral vector; and T. Kitamura for the pMY vector and PlatE cells. Y.S. is a fellow of the Susan G. Komen Breast Cancer Foundation. D.C.L. was supported by the Deutsche Forschungsgemeinschaft. P.T. was supported by the Pasarow Foundation. K.N. was supported by a Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Research Abroad. R.M.E. is an Investigator of the Howard Hughes Medical Institute at the Salk Institute and March of Dimes Chair in Molecular and Developmental Biology. F.H.G. is the Adler Professor of Age-Related Neurodegenerative Diseases. This work was supported by the Howard Hughes Medical Institute, the NIH, the Christopher Reeve Paralysis Foundation, the National Institutes of Aging, the Michael J. Fox Foundation, Project ALS and the Lookout Fund.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M.E. (evans@salk.edu) or F.H.G. (gage@salk.edu).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Spain/

Portugal/ Belgium:

Evelina Rubio Hakansson (4973)

Scandinavia: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
Asia-Pacific Sales Director:
Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

On the road

According to the Chinese calendar, 2004 is the year of the monkey, but for Europe it looks more likely to be the year of mobility. That, at least, was one of the key points in a talk given by Raffaele Liberali last month at the New York Academy of Sciences. Liberali, who heads the division for mobility and Marie Curie fellowships at the European Commission's research directorate, said that he wants to fix two major problems that have left Europe trailing in the wake of the United States as a destination of choice for researchers (see www.nyas.org).

First, Liberali said that he will try to make the commission's grants more portable between countries so that researchers can carry funding from, say, a university in Spain to one in the United Kingdom. In the United States, it is common for researchers to take National Institutes of Health or National Science Foundation grants with them when they change jobs. Second, Liberali said that he would try to establish a code of conduct for the recruitment of researchers. Some European universities, he noted, are not transparent in their hiring process, which results in patronage and can hamper researchers who are 'outside' the local system.

The European Commission plans to use a series of grants to help it accomplish its goals. Some of its grants, such as the Marie Curie postdoc fellowships, only fund researchers from one country to work in another, thus encouraging mobility. The commission is also increasing the number of 'reintegration' grants to help researchers who work for a time in the United States to return to Europe — although it can't guarantee that these will automatically allow researchers to return to their home country.

Such schemes mean that, if you are a young European researcher who is having trouble getting a grant in your home country, you might be better off riding the European mobility funding stream, rather than trying to stay put.

Paul Smaglik

Naturejobs editor



WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS